Ensemble auf dem Ruinenberg

Echoes of Interdisciplinary Research with Angela Sasse

Aad van Moorsel

School of Computer Science
University of Birmingham, Birmingham, UK
a.vanmoorsel@bham.ac.uk

Abstract. This short contribution traverses through a number of professional and scientific interconnections that popped up when reflecting on the research projects I worked on alongside Angela Sasse. These projects exhibited a unique blend of interdisciplinary research that combined computer science and mathematical modeling with insights from social sciences and human psychology. This free-spirited interdisciplinarity resonates widely in modern-day security and trustworthiness research dominated by the proliferation of AI. In this paper, I use ensembles to reflect on challenges to building trustworthy systems in the era of AI. AI ensembles, with its fascinating connections to economics, political sciences and trust, reminded me strongly of the inspiring interdisciplinary discussions we had in the Research Institute in the Science of Cyber Security that Angela led. In addition, purely coincidentally, I am currently the Head of the School of Computer Science at the University of Birmingham, where Angela received her PhD. I took the opportunity to reexamine if the dissertation was a premonition of the career that followed (as I will explain, I believe the answer to be both yes and no).

1 Introduction

The day of the deadline for submission to the Festschrift, I undertook a 'Spaziergang' to the *Ensemble auf dem Ruinenberg* in Potsdam, near Berlin, Germany. The pictures in Fig. 1 stem from that family trip, the left showing some of the ruins on the top of the hill that were built for Frederick the Great in the 18th century. The sign on the right of Fig. 1 explains that the ruins served as an architectural framing of the water reservoir at the top of the hill, which irrigated the surrounding lands.

When visiting the site, the Ensemble auf dem Ruinenberg struck me as a metaphor for traditional computer science disciplines, which often seem dated

Fig. 1. Das Ensemble auf dem Ruinenburg: A metaphor for traditional science and engineering disciplines in the times of AI? (Photographs from the author.)

ruins in an era dominated by astounding progress in AI. Traditional disciplines such as software engineering, dependability and HCI face questions about their relevance when AI can do the same job, and when technology progresses at a pace that almost prohibits careful engineering. Scientific progress along the lines of user-centred design struggles with the pace of AI progress, and traditional design disciplines such as dependability and security are challenged by the fact that AI is not only the problem, but also the proposed solution: if AI leads to challenges, eg in terms of security or reliability, resolve it by throwing more AI at it. At the end of the paper I come back to this point, and argue that what is really needed is an integrated and possibly unified understanding of AI and traditional computer science.

The societal impact of AI in the current era has led to a proliferation of interdisciplinary concerns that has similarities with research in past projects with Angela Sasse. I have had the pleasure to work alongside Angela in two large research projects while I was at Newcastle University. The Trust Economics project was led by David Pym and colleagues at my prior employer Hewlett Packard, advocating discrete-event system modeling to quantify security improvements, exemplified by [2,5] and applied successfully in industry setting. This work sew the seeds for UCL's REF impact case in 2021 [20], based on Angela's research with Adam Beautement and Simon Parkin (who was working with me in Newcastle during the Trust Economics project). The second project collaboration was through our respective projects within the EPSRC/NCSC Institute for Sociotechnical Cyber Security, for which Angela Sasse was the founding Director. For three years the Newcastle/Northumbria team of myself, Pam Briggs, Lynn Coventry and postdocs James Turland and Debora Jeske took regular trips on the 7am Flying Scotsman train to join the other project teams at our meetings at UCL in London. A great time was had by all!

One of the most striking AI-related examples of hidden connections between different disciplines is the *ensemble* approach to improving the accuracy of AI-

based services. In a recent paper [14] I have been discussing the similarities between fault tolerance techniques developed to improve the dependability of traditional computer systems, and AI techniques that improve the accuracy of the AI service. There is an almost one-to-one mapping between dependability architectural concepts and (post-learning) architectural approaches to improve AI's accuracy, with the ensemble as the most intriguing. Since I visited the Ensemble auf dem Ruinenberg at the day of the paper deadline, writing about ensembles seems the only appropriate manner to reflect on joint interdisciplinary research in Angela Sasse's Festschrift.

Another coincidence that fed into this paper is that I am the current Head of the Department where Angela received her PhD in Computer Science, at the University of Birmingham in the UK. Angela's PhD [17,18] considered research methods for HCI anchored in social sciences. Such methods are still not broadly appreciated across computer science, and Angela and supervisor Dennis Parkyn must have been trail blazers in the nineties. I was not able to obtain a first-hand account from people associated with Angela's research in Birmingham, but reading the dissertation, almost 30 years after publication, it is still extremely readable and relevant. It is a defense of science, with a very fresh and open-minded approach to how high-quality science in HCI should develop. More about the dissertation in Section 2.

Angela Sasse's research in usable security has been exceptionally impactful, as the 2021 REF impact case document evidences [20]. Of course, the academic outcomes and societal impact cannot be achieved by a single person, and benefited strongly from the talents of a variety of exceptional researchers, some of which I mentioned above, who helped drive the research and co-authored the papers. The resulting research exhibits exceptional rigour and demonstrates a rare ability to articulate the bottom line implications, in a manner that is meaningful in equal measure to academics and practitioners. The developed theories and insights in usable security are unparalleled in profound depth and diversity, including those in compliance budget [3], security by stealth [15], shadow security [9], etc. Not to mention the extraordinary talent in communicating using titles of papers, The User is Not the Enemy [1] the most prominent but far from only highly effective title. The careful reader may have observed an attempt to achieve similar results in this paper, but unfortunately it lacks the gift of the pen that Angela's papers demonstrated again and again.

2 Angela Sasse's PhD at University of Birmingham

As Angela is an alumna of the School I currently lead, I was curious to find out if Angela Sasse's PhD dissertation can be regarded as a premonition of the groundbreaking work in human aspects of cyber security that would follow. I found that the answer is both a profound 'yes' (the PhD sets the scientific foundations for the later work) and and obvious 'no' (the PhD did not address cyber security).

The dissertation has two parts, both thick and heavy volumes when printed in the nineties. Part one is the 200-page dissertation [17], while the second part is an addendum with transcripts of the five case studies conducted [18]. The dissertation studies mental models of users of technology, and particularly investigates and innovates in appropriate research methods for Human-Computer Interaction.

The Research Discipline of HCI. The hidden gem of the dissertation, in my view, is the discussion about The Research Discipline of HCI, Chapter 2. Written in the mid nineties, it very eloquently sets out and discusses the emerging approaches to HCI research, from pure (quantitative) science, to design and as an engineering discipline. These approaches are core to the debate at that time about how to make scientific progress given the inherently interdisciplinary nature of HCI, and the discussion provides perspectives on value and needs of hard and soft science. The chapter notes shortcomings with all three, and most clearly distances itself from over-reliance on the 'hard' science approach, correctly pointing out that computer science in general does not adhere to a highly scientific quantitative approach to begin with. Presented almost as an afterthought, grounded theory is suggested as a promising method of choice: "the HCI research community should consider treating HCI knowledge as data from which the conception for the discipline can be generated by using grounded theory."

In addition to this remark about grounded theory, the conclusions section of the dissertation provide plenty of hints of the direction Angela's research would take. It states that "researchers keep producing conceptualisations of users' models, rather than examining users' models themselves." The dissertation advocates constructive interaction scenarios, to allow users to communicate their thoughts and knowledge. The conclusion is that "HCI researchers wishing to support designers have to abandon their current compartmentalised model of these processes, identify mechanisms which can influence the model construction process, and specify how they influence the way in which models are used."

No Cyber. Reading the dissertation it becomes clear quickly that it is highly ambitious, extremely well written, but has little or no connection to cyber security. Angela had started working in this period with Anne Adams, leading to the 1999 paper 'Users Are Not the Enemy' [1], but in that period also frequently publishes in other technology areas, e.g., multimedia quality of service.

I would be tempted to interpret the dissertation as a 'defense of science', as it argues against dogmatism and fervently in favour of methods that establish a true, deep, understanding of users and users' mental models. Its ambitious aim is to establish appropriate research methods for the HCI community, in support of research that is timely and thus relevant, and, most importantly, in support of research that is rigorous, honest and unbiased. Arguably, the methods developed and studied in the dissertation are the staple of Angela's research throughout her career. Although the dissertation does not provide indications for the unique impact Angela Sasse would go on to have in cyber security research and prac-

tice, the dissertation presents the development of the underlying research skills, establishing the foundations required for the later research.

3 Ensembles

We return to the main story line, that of ensembles. In all of Angela Sasse's research, she demonstrated a rare ability to link IT security concerns to research across disciplinary boundaries. She therefore was uniquely well positioned to be the founding Director of the RISCS Research Institute and to oversee four different projects, each with its own interdisciplinary mixture. The Newcastle team worked on choice architecture (nudging in popular terminology), developing the SCENE methodology for design of systems using nudging [6] and IT security applications of this methodology [19]. Identical to nudging, the motivation for AI ensembles comes from economics and social sciences, for ensembles through the Jury Theorem by Condorcet from 1785 [13].

AI Ensembles. Ensembles run in parallel a number of AI services for the same task, and then conduct a majority vote to decide the outcome. For example, assume an AI classifier that distinguishes between cats and dogs, or recognises stop signs in autonomous vehicles. One then sets up a number of different classifiers, which differ in the data used for learning or in the specific algorithm used. Each input image is classified by each of the classifiers, which deliver their outcome to a software-implemented voter, where the majority decides.

Jury Theorem. It is not necessarily intuitively obvious whether an AI ensemble improves the accuracy: how can a number of inferior AI algorithms do better than the one algorithm that is known to be the best? The core argument in favour of ensembles was provided two and a half century ago by Condorcet's extremely impressive essay [13]. Condorcet provided us with the Jury Theorem as well as with the underpinning probabilistic reasoning. The Jury Theorem essentially explains why a jury, under certain conditions, can be shown to be preferred over a single judge, despite the poorer judgment of each individual juror.

Fig. 2 demonstrates the dramatic improvement one achieves with an ensemble using a majority decision. The model assumptions behind the improvements in Fig. 2 is that all classifiers have the same independent probability to be correct (provided on the x-axis). If the probability of correctness is higher than 0.5, the majority vote becomes better with increasing numbers of AI algorithms. The number of classifiers does not need to be very high, already with a modest number of classifiers the ensemble works much better. Note also that in the context of human decision making such as parliamentary voting, if individuals are correct less than probability 0.5 the outcome is disastrous, and the wrong decision is taken as majority vote.

An excellent survey of the methods and basic approaches for AI ensembles can be found in Rokach's survey from 2010 [16]. Hansen [7] in 1990 provided basic analytics to support and explain the working of ensembles, and Kucheva [10] provides grounding for diversity measures for ensembles. The latter work is

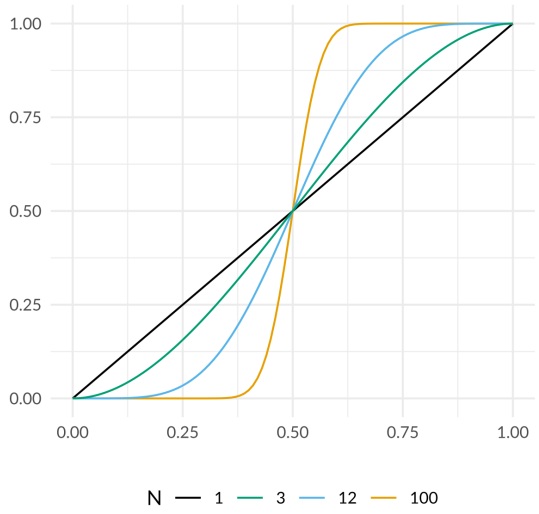

Fig. 2. An ensemble with N classifiers. The y-axis shows the probability that the ensemble outcome is correct after majority voting; the x-axis shows the probability that individual classifiers are correct. Source: Condorcet's Jury Theorem [8], CC BY-SA 4.0. No changes made.

very interesting, the argument for the practical success of ensembles is not in independence, but in diversity of AI algorithms. In fact, one can purposely introduce negative correlations between algorithms, e.g., through the used training data, to improve the outcome of the ensemble.

N-Modular Redundancy. In dependability, one of the most standard manners of providing fault tolerance is that of adding redundant copies, and then vote to establish the outcome. If there are N copies, this is called N-Modular Redundancy (NMR), and with 3 copies, it is called triple modular redundancy. When the modules are software modules, the method is termed N-Version Programming. In NMR, a voter decides about the outcome, typically based on majority voting. As long as a majority of the modules provides the correct outcome, the outcome of NMR is correct. As reference for the work on software fault tolerance, we refer to the 1995 book *Software Fault Tolerance* edited by Michael Lyu [12], which contains chapters on all fault tolerance mechanisms mentioned here.

Contrary to AI ensembles, one can argue that it is intuitively obvious that NMR is likely to improve the situation. If it is rare that a component fails, then with three or more copies, it will be almost impossible for the majority to fail. In hardware fault tolerance, where it is highly unpredictable how faults will manifest themselves, this type of redundancy is highly effective. For software, it is harder to establish effective modular redundancy, since either the software or the inputs must be different from each other (the latter not always being possible).

To produce multiple copies of software is usually prohibitively expensive, and even if different teams code different pieces of software, the faults may manifest themselves in the same 'tough' parts of the code, failing on the same inputs. In that case, redundant copies do not help, they all fail in the same way.

Although there are important differences between ensembles and N-Modular Redundancy, the key to both is *diversity*, and in both fields the analysis of diversity is critical. For AI ensembles, the work by Kucheva [10] provides a systematic study of diversity, while for dependability, Littlewood and Miller [11] provide a deep system analysis of N-Version Programming and its requirements for diversity.

4 Discussion

The point of the above discussion and of the recent exploration of dependability in the time of AI [14] is to debate the uneasy marriage of AI and dependability or trustworthiness. (See also [4,21] for perspectives stemming from the dependability community.) On the one hand, the uncertain outcomes of AI make one fearful to use it in applications that need to be dependable. After all, there are fundamental limits to the accuracy one can achieve by adding training data and to the ability to control the generalisation error. On the other hand, Fig. 2 shows there are approaches to improve accuracy, theoretically to arbitrary level. Through fault tolerance approaches such as advocated in [4], with AI guarding AI, one can establish highly dependable solutions, so even though AI introduces uncertainty, this does not need to imply it needs to be subject to more failures.

The working of AI ensembles can be explained from economics and social sciences, and the technique is directly related to N-Modular Redundancy in dependability. Despite the similarities between AI and dependability architectures, we still lack integrated understanding to bridge between the statistical nature of AI accuracy and the deterministic nature of removing, tolerating and managing faults. Returning to the question we started with, namely whether traditional computer science approaches are essentially 'ruins' in the era of AI, I would argue the answer is a definite 'No'. It will be essential to arrive at an integrated understanding of AI and traditional system architecture approaches, to be able to benchmark and evidence our reliance on AI-based systems. The definition of dependability is 'justified reliance', building sound AI-based systems is not only about avoiding failures, it is also about understanding these systems well enough to *justify* the trust we place on these systems.

References

1. Adams, A., Sasse, M.A.: Users are not the enemy–why users compromise computer security mechanisms and how to take remedial measures. Communications of the ACM **42**(12), 40–46 (1999)

2. Beautement, A., Coles, R., Griffin, J., Monahan, B., Pym, D., Sasse, M.A., Wonham, M.: Modelling the human and technological costs and benefits of USB memory stick security. In: Proceedings of the Workshop on Economics in Information Security (WEIS 2008) (2008)
3. Beautement, A., Sasse, M.A., Wonham, M.: The compliance budget: Managing security behaviour in organisations. In: Proceedings of the 2008 New Security Paradigms Workshop (NSPW 2008). pp. 47–58 (2008)
4. Bloomfield, R., Rushby, J.: Assurance of AI systems from a dependability perspective. Tech. Rep. SRI-CSL-2024-02, Computer Science Laboratory, SRI International, Menlo Park, CA (July 2024), https://www.csl.sri.com/~rushby/abstracts/aisafety24, updated June 2025; also available on arXiv:2407.13948
5. Coles, R., Griffin, J., Johnson, H., Monahan, B., Parkin, S.E., Pym, D., Sasse, M.A., van Moorsel, A.: Trust economics feasibility study. In: Proceedings of the 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2008). pp. A45–A50. IEEE Computer Society (2008)
6. Coventry, L., Briggs, P., Jeske, D., van Moorsel, A.: SCENE: A structured means for creating and evaluating behavioral nudges in a cyber security environment. In: Proceedings of the International Conference on Design, User Experience, and Usability. pp. 229–239 (2014)
7. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(10), 993–1001 (1990). https://doi.org/10.1109/34.58871
8. Jaszczuroczlek: Condorcet's jury theorem graph.svg (2026), https://en.wikipedia.org/wiki/Condorcet%27s_jury_theorem, wikipedia article, licensed under CC BY-SA 4.0
9. Kirlappos, I., Parkin, S., Sasse, M.A.: Shadow security as a tool for the learning organization. In: Proceedings of the New Security Paradigms Workshop (NSPW 2014). pp. 29–38. ACM SIGSAC, ACM (2014)
10. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles. Machine Learning **51**(2), 181–207 (2003). https://doi.org/10.1023/A:1022859003006
11. Littlewood, B., Miller, D.R.: Conceptual modelling of coincident failures in multiversion software. IEEE Transactions on Software Engineering **15**(12), 1596–1614 (1989). https://doi.org/10.1109/32.58744
12. Lyu, M.R. (ed.): Software Fault Tolerance. Wiley Trends in Software, John Wiley & Sons, Chichester, UK (1995)
13. Marquis de Condorcet: Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris (1785), https://gallica.bnf.fr/ark:/12148/bpt6k417181
14. Moorsel, A.v.: Assessing fault tolerance architectures for AI classifiers. In: Thomas, N. (ed.) Collection of Articles from UK Performance Engineering Workshop (2025)
15. Parkin, S., van Moorsel, A., Inglesant, P., Sasse, M.A.: A stealth approach to usable security: Helping IT security managers to identify workable security solutions. In: Proceedings of the New Security Paradigms Workshop (NSPW) (2010)
16. Rokach, L.: Ensemble learning: A survey. Pattern Recognition **43**(9), 293–304 (2010). https://doi.org/10.1016/j.patcog.2010.04.001
17. Sasse, M.A.: Eliciting and Describing Users' Models of Computer Systems. Ph.D. thesis, University of Birmingham, Birmingham, UK (April 1997)
18. Sasse, M.A.: Eliciting and Describing Users' Models of Computer Systems–Volume 2 Transcripts. Ph.D. thesis, University of Birmingham, Birmingham, UK (April 1997)

19. Turland, J., Coventry, L., Jeske, D., Briggs, P., van Moorsel, A.: Nudging towards security: Developing an application for wireless network selection for Android phones. In: Proceedings of the 2015 British HCI Conference. pp. 193–201 (2015)
20. UCL: Human-centred security policy. Research Excellence Framework 2021 Impact Case Study (2021), https://results2021.ref.ac.uk/impact/188a46c6-048b-4da1-bd4b-adf8321cbdb4
21. Vieira, M.: Why we should trust systems, not just their AI/ML components. Computer **58**(11), 84–94 (2025)