Report from Dagstuhl Seminar 21231

Transparency by Design

Edited by

Judy Kay[1], Tsvi Kuflik[2], and Michael Rovatsos[3]

1 **The University of Sydney, AU,** judy.kay@sydney.edu.au
2 **Haifa University, IL,** tsvikak@is.haifa.ac.il
3 **University of Edinburgh, GB,** michael.rovatsos@ed.ac.uk

Abstract

This report documents the program and outcomes of Dagstuhl Seminar 21231 on "Transparency by Design" held in June 2021. Despite extensive ongoing discussions surrounding fairness, accountability, and transparency in the context of ethical issues around AI systems that are having an increasing impact on society, the notion of transparency – closely linked to explainability and interpretability – has largely eluded systematic treatment within computer science to date.

The purpose of this Dagstuhl Seminar was to initiate a debate around theoretical foundations and practical methodologies around transparency in data-driven AI systems, with the overall aim of laying the foundations for a "transparency by design" framework – a framework for systems development methodology that integrates transparency in all stages of the software development process. Addressing this long-term challenge requires bringing together researchers from Artificial Intelligence, Human-Computer Interaction, and Software Engineering, as well as ethics specialists from the humanities and social sciences, which was a key objective for the four-day seminar conducted online.

Seminar June 6–10, 2021 – http://www.dagstuhl.de/21231
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI); Software and its engineering → Software development techniques
Keywords and phrases Artificial Intelligence, Dagstuhl Seminar, Ethics, Human-ComputerInteraction, Software Engineering, Transparency
Digital Object Identifier 10.4230/DagRep.11.5.1
Edited in cooperation with Bogina, Veronika, Orphanou, Kalia, Podoletz, Lena, Shulner-Tal, Avital

1 Executive Summary

Judy Kay (The University of Sydney, AU)
Tsvi Kuflik (Haifa University, IL)
Michael Rovatsos (University of Edinburgh, GB)

As AI technologies are witnessing impressive advances and becoming increasingly widely adopted in real-world domains, the debate around the ethical implications of AI has gained significant momentum over the last few years. Much of this debate has focused on fairness, accountability, transparency and ethics, giving rise to "Fairness, Accountability and Transparency" (FAT or FAccT) being commonly used to capture this complex of properties as key elements to ethical AI.

Transparency by Design, *Dagstuhl Reports*, Vol. 11, Issue 05, pp. 1–22
Editors: Judy Kay, Tsvi Kuflik, and Michael Rovatsos
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

However, the notion of transparency – closely linked to terms like explainability, accountability, and interpretability – has not yet been given a holistic treatment within computer science. Despite the fact that it is a prerequisite to instilling trust in AI technologies, there is a gap in understanding around how to create systems with the required transparency, from demands on capturing their transparency requirements all the way through to concrete design and implementation methodologies. When it comes to, for example, demonstrating that a system is fair or accountable, we lack usable theoretical frameworks for transparency. More generally, there are no general practical methodologies for the design of transparent systems.

The purpose of this Dagstuhl Seminar was to initiate a debate around theoretical foundations and practical methodologies with the overall aim of laying the foundations for a "Transparency by Design" framework, i.e. a framework for systems development that integrates transparency in all stages of the software development process.

To address this challenge, we brought together researchers with expertise in Artificial Intelligence, Human-Computer Interaction, and Software Engineering, but also considered it essential to invite experts from the humanities, law and social sciences, which would bring an interdisciplinary dimension to the seminar to investigate the cognitive, social, and legal aspects of transparency.

As a consequence of the Covid-19 pandemic, the seminar had to be carried out in a virtual, online format. To accommodate the time zones of participants from different parts of the world, two three-hour sessions were scheduled each day, with participant groups of roughly equal size re-shuffled each day to provide every attendee with opportunities to interact with all other participants whenever time difference between their locations made this possible in principle. Each session consisted of plenary talks and discussion as well as work in small groups, with discussions and outcomes captured in shared documents that were edited jointly by the groups attending different sessions each day.

The seminar was planned to gradually progress from building a shared understanding of the problem space among participants on the first day, to mapping out the state of the art and identifying gaps in their respective areas of expertise on the second day and third day.

To do this, the groups identified questions that stakeholders in different domains may need to be able to answer in a transparent systems, where we relied on participants to choose domains they are familiar with and consider important. To identify the state of the art in these areas, the group sessions on the second and third days were devoted to mapping out the current practice and research, identifying gaps that need to be addressed.

The two sessions on each day considered these in terms of four aspects: data collection techniques, software development methodologies, AI techniques and user interfaces.

Finally, the last day was dedicated to consolidating the results towards creating a framework for designing transparent systems. This began with each of the parallel groups considering different aspects: Motivating why transparency is important; challenges posed by current algorithmic systems; transparency-enhancing technologies; a transparency by design methodology; and, finally, the road ahead.

The work that began with the small group discussions and summaries continued with follow up meetings to continue the work of each group. The organisers have led the work to integrate all of these into an ongoing effort after the seminar, aiming to create a future joint publication.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 The Limits of Transparency

Joanna J. Bryson (Hertie School of Governance – Berlin, DE)

Main reference Joanna J. Bryson: "The Artificial Intelligence of the Ethics of Artificial Intelligence", The Oxford Handbook of Ethics of AI, p. 1, Oxford University Press, 2020.
URL https://doi.org/10.1093/oxfordhb/9780190067397.013.1

Among many strong and positive suggestions in the 2020 EU whitepaper on AI was at least one repeated falsehood: that AI is necessarily opaque. AI is of necessity no more opaque than natural intelligence; in fact, digital artifacts can by choice be made to be far more transparent than nature. In the talk, I describe technological, sociological, and economic barriers to transparency, how these are affected by AI and the digital revolution, and what governance policies may be deployed to address them. Here in this extended abstract, I just talk about what transparency means, and what it is for.

My intent here is to provide the definitions most useful for interpreting the five OECD Principles of AI, since that's the soft law with the most international support, with 50 national governments (the OECD and the G20) now signed up to it.

Responsibility is the keystone. It is a property assigned by a society to individuals for their actions, including inactions where action was the expected norm. Actors with responsibility are technically termed 'moral agents' in philosophy. These vary by society. So for example, a family may hold a child or a dog responsible to know where an appropriate place is to pee. But a government will only hold legal persons responsible. In the case of a family, those will be the adult humans in the household.

Accountability is the capacity to assign responsibility to the correct agency. The purpose of accountability and responsibility are to maintain social order, that is, to maintain the society. Therefore responsibility is ordinarily assigned to those who can be held to account. The purpose of holding people to account is to persuade them and others like them to perform what a society considers to be their responsibilities for maintaining that society. Sometimes responsibility is determined to fall outside the control of any agency. In the USA for example, weather events are frequently termed "acts of God."

Transparency is the property of a system whereby it is possible to trace accountability and allocate responsibility. Where there is transparency, there does not need to be blind trust. Formally, trust is the expectation of good behaviour afforded (by a truster) to others (trustees) where the trustee's behaviour is actually unknown and where the trustee actually has autonomy with respect to the truster. Transparency may create a sensation of trust, but it renders the formal state of trust unnecessary. Transparency is something that can and should be designed into an intelligent system.

I postulate in my talk that the limits of transparency are problems like computational combinatorics, which means we can never know everything in detail; sociological problems like political polarisation and identity politics, which make people not likely to believe information even if it is in front of them; and mutually exclusive goals, so for example there can be no meeting of the minds when one mind is focussed only on wealth creation and another mind is focussed only on ethics red lines. But AI does not make any of these problems worse. The fact that AI systems contain complicated components is no more of a problem than organisations composed of humans with complicated brains. We only need transparency as to who is responsible for ensuring a system has been developed and operated in such a way that if it functions incorrectly or to malign purpose, we can know who caused it to do such a thing, even if that cause is failure to follow best practice or the release of an inadequately tested or understood system.

3.2 Fairness-Aware Recommender Systems

Robin Burke (University of Colorado – Boulder, US)

Recommender systems are machine learning systems that provide personalized results to users across a wide array of applications from social media to e-commerce to news to online dating. As fairness in machine learning has become a major sub-field of research in the past five years, recommender systems have also benefited from this emphasis. However, as these fairness-aware systems begin to be deployed, it becomes quite clear that we know very little about how users think about and interact with systems that take ethical stances, stances which might put them at odds with user interests and goals. The multi-sided nature of recommender systems also becomes clear when we consider fairness and this suggests the need for transparency toward providers of items being recommended. Fairness-aware recommendation therefore raises two key challenges: (1) how best to implement transparency for users who consume recommendations designed to be fair and (2) how to implement transparency for item providers for whom fairness may be important but where transparency may enable adversarial manipulation by such users.

3.3 Artificial Intelligence for Social Good: When Machines Learn Human-like Biases from Data

Aylin Caliskan (University of Washington – Seattle, US)

Developing machine learning methods theoretically grounded in implicit social cognition reveals that unsupervised machine learning captures associations, including human-like biases, objective facts, and historical information, from the hidden patterns in datasets. Machines that learn representations of language from corpora embed biases reflected in the statistical regularities of language. Similarly, image representations in computer vision contain biases due to stereotypical portrayals in vision datasets. On the one hand, principled methodologies for measuring associations in artificial intelligence provide a systematic approach to study society, language, vision, and learning. On the other hand, these methods reveal the potentially harmful biases in artificial intelligence applications built on general-purpose representations. As algorithms are accelerating consequential decision-making processes ranging from employment and university admissions to law enforcement and content moderation, open problems remain regarding the propagation and mitigation of biases in the expanding machine learning pipeline.

3.4 Towards Personalized Explainable AI

Cristina Conati (University of British Columbia – Vancouver, CA)

The AI community is increasingly interested in understanding how to build artifacts that are accepted and trusted by their us-ers in addition to performing useful tasks. It is undeniable that explainability can be an important factor for acceptance and trust. However, there is still limited understanding of the actual relationship between explainability, acceptance, and trust and which factors might impact this relationship. In particular, although existing research on Explainable AI (XAI) suggests that having AI systems explain their inner workings to their end users can help foster transparency, interpretability, and trust. there are also results suggesting that such explanations are not always wanted by or beneficial for all users. These results indicate that research in XAI needs to go beyond one-size-fits-all explanations and investigate AI systems that can personalize explanations of their behaviors to the user's specific needs.

There is general agreement that such needs may depend on context, e.g., the type of AI application and criticality of the targeted tasks, but there is also evidence that, given the same context, *user differences* play a role in defining when and how explanations may be useful and effective.

These results call for the need to investigate *personalized XAI*, namely how to create AI systems that understand to whom, when and how to deliver effective explanations of their actions and decisions.

AI-driven personalization has been an active field of research for several decades, spanning fields such as recommender systems, intelligent-tutoring systems, conversational agents, and affect-aware systems. To provide personalization, an AI system needs to have an adaptive loop in which it acquires a model of its user by inferring relevant user properties from available observations and decides how to personalize its behavior accordingly, to favor at best the goal of the interaction

We see explanations as yet another element of personalization in the adaptive loop, where the system ascertains if and how to justify its behavior to the user based on its best understanding of user properties specifically relevant to evaluate the need for explanation. What these relevant properties are is still largely unknown, hence, a key step toward personalized XAI is re-search to fill this gap.

Two general types of user properties have shown to be relevant for personalization: long-term traits that do not usually change over short periods of time (such as cognitive abilities and personality traits); and transient short-term states such as attention, interest and emotions.

We argue that, given a specific AI application, different types and forms of explanations may work best for different users, and even for the same user at different times, depending to some extent on both their *long-term traits* and *short-term states.* As such, our long-term goal is to develop personalized XAI tools that adapt dynamically to the user's needs by taking both these types of user factors into account.

In this talk, I focus on research investigating the impact of *long-term traits*, and how they may drive personalization. I present a general methodology to address these two questions, followed by an examples of how it was applied to gain insights on which long-term traits are relevant for personalizing explanations in an intelligent tutoring system (ITS). I discuss how to move forward from these insights, and present research paths that should be explored to make personalized XAI happen.

References

1 C. Conati, O. Barral, V. Putnam, and L. Rieger. Toward personalized XAI: A case study in intelligent tutoring systems, *Artificial Intelligence*, 298, 2021.

3.5 Increasing Transparency with Humans in the Loop

Gianluca Demartini (The University of Queensland – Brisbane, AU)

Bias appears in data collected from human annotators. It is then propagated into the Artificial Intelligence (AI) models trained with such labelled datasets. Bias in AI is then presented to end users who interact with the AI-powered system with their own bias and stereotypes. In such a setting, increasing the level of transparency could be an alternative to popular approaches aimed instead at removing bias from the system.

In this talk, I first present an example from our recent research of bias present in labelled datasets generated by human annotators in the context of crowdsourced judgements of information truthfulness. I then discuss how data-driven models of human annotator interaction behaviour may be leveraged to better understand the behavioural diversity present in a group of human annotators and the potential bias reflected in the labels generated by them. Finally, I discuss possible approaches to manage such bias in data and AI going beyond the classic aim of removing it.

References

1 La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., and Spina, D. (2020, April). Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In European Conference on Information Retrieval (pp. 207-214). Springer, Cham.

2 Han, L., Roitero, K., Gadiraju, U., Sarasua, C., Checco, A., Maddalena, E., and Demartini, G. (2019, January). All those wasted hours: On task abandonment in crowdsourcing. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 321-329).

3 JHan, L., Checco, A., Difallah, D., Demartini, G., and Sadiq, S. (2020, October). Modelling User Behavior Dynamics with Embeddings. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 445-454).

3.6 Transparency by design

Virginia Dignum (University of Umeå, SE)

In this talk, discuss the desirability and challenges of a design approach to transparency and proposed it to be complemented and extended by a "transparency in design" approach that focus on the processes, choices and stakeholders. Current work on transparency by design focuses on algorithmic transparency, namely on the data, results and functionalities of algorithms but often ignores the context in which algorithms are developed and used, and the power relations that influence decisions and requirements. Moreover, algorithmic transparency is not always possible nor desirable (IP, security, complexity, etc) so it needs to be complemented by methods that build trust, such as contracts and formal governance processes.

3.7 Transparency issues for the Use of AI and Analytics in Financial Services

Ansgar Koene (EY Global – London, GB)

Data analytics, risk modelling and prediction are common practices with a long tradition in the financial services. Despite great interest and a large number of pilot projects to explore the use of AI for enhancing the power of these computational practices, however, established financial institutions have been slow to put AI into operational use for core financial activities. Key factors that are holding back the deployment of AI are concerns about over the ability to comply with regulatory requirements regarding the explainability and robustness of financial models.

In this talk I review existing regulatory requirements and compliance considerations for the use of data analytics in the financial industry, which must be considered when evaluating the potential for the use of AI in this sector. Starting with an overview of levels of governance requirements that apply to the use of models in financial services and an associated algorithm risk tiering I provide some examples of typical model governance approaches for different types of regulated model classes. Based on these established model governance requirements I consider four dimensions of risk associated with AI: the data; the models and their implementation; the approach to modelling; and the (lack of) accumulated experience for these types of models. When considering the mitigation of these risks, I briefly present some model validation and monitoring considerations for conceptual soundness, data quality review, outcomes analysis, implementation controls and performance monitoring that are associated with business knowledge, data governance, cross validation, IT governance and model after-care, respectively.

Zooming in on model explain ability, I consider the familiar issue of trade-off between model interpretability and accuracy by focusing in on additional objectives that interpretability serves, such as: debugging and improvement of the model; trust and acceptance; regulatory compliance; ethics; safety and transferability; and discovery of unknown relationships in the data by being better able to interpret model outcomes. Particular attention is paid to the challenge of addressing the different explainability needs of various stakeholders, as highlighted in the Bank of England working paper No816 on "Machine learning explainability in finance" an application to default risk analysis.

I conclude with some findings of a survey on the ways in AI is being used by banks, or how they anticipate using it in the near future.

3.8 Engineering Traceability: A Lens Connecting Transparency Tools to Accountability Needs

Joshua A. Kroll (Naval Postgraduate School – Monterey, US)

Accountability is widely understood as a goal for well governed computer systems, and is a sought-after value in many governance contexts. But how can it be achieved? Many authors suggest it is enabled by transparency, though without a clear mechanism or requirements this, too, is challenging to achieve adequately. Recent work on standards for governable artificial intelligence systems offers a related principle: traceability. Traceability requires establishing not only how a system worked but how it was created and for what purpose, in a way that explains why a system has particular dynamics or behaviors. It connects records of how the system was constructed and what the system did mechanically to the broader goals of governance, in a way that highlights human understanding of that mechanical operation and the decision processes underlying it. We examine the ways that traceability links transparency demands to accountability needs, distill from these a set of requirements on software systems driven by the principle, and systematize the technologies available to meet those requirements. From our map of requirements to supporting tools, techniques, and procedures, we identify gaps and needs separating what traceability requires from the toolbox available for practitioners. This map reframes existing discussions around accountability and transparency, using the principle of traceability to show how, when, and why transparency can be deployed to serve accountability goals and thereby improve the normative fidelity of systems and their development processes.

3.9 Transparency and the Fourth AI Revolution

Loizos Michael (Open University of Cyprus – Nicosia, CY)

Facilitated by the desire to scientifically understand and replicate human intelligence in machines, the First AI Revolution had as a primary consequent the offloading of the cognitive burden of humans – reminiscent of the offloading of the physical burden of humans during the First Industrial Revolution – with humans retaining control as domain experts in their interaction with machines. This central role of humans was considerably diminished during the Second AI Revolution, where the advent of Deep Learning drove humans to the subsidiary role of "blue-collar" workers annotating data over the machine learning "assembly line" – echoing the primary consequent of the Second Industrial Revolution – and raised concerns on humans ceding too much control to AI.

The ongoing effort towards building transparent AI can, ultimately, be seen as a natural reaction and a potential remedy to these concerns. In this context, post-hoc transparency on why a system exhibits a certain behavior is not sufficient. Rather, transparency mechanisms should be built by design into an AI system to reveal why the latter's exhibited behavior is what it should be, according to any applicable legal, social, and ethical frameworks. The direction foreshadowed by the need for such transparency is that of an oncoming Fourth AI Revolution, facilitated – as with the Fourth Industrial Revolution – by an increased communication between humans and machines, and a resulting transition from machine automation to machine autonomy that is guided, monitored, and evaluated by humans, towards building long-term trust and offering reassurances that machines will respect the values that humans deem important.

In this talk we argue that this form of transparency can be achieved without abandoning the central role that machine learning has played in AI so far, but by extending the role of humans from that of data annotators to that of machine coaches. Much like how humans help each other acquire new knowledge by interactively providing feedback, the Machine Coaching paradigm that we put forward proceeds on the basis that humans and machines engage in a dialog on why a certain behavior was exhibited by a machine: the machine provides an explanation based on its current knowledge, and if the human finds the explanation lacking then the latter effectively provides back an improved or more appropriate explanation that the machine proceeds to integrate into its learned knowledge for future use. Feedback bilaterally exchanged between humans and machines in the context of Machine Coaching is, therefore, specific to the given situation, and is provided in-situ and in an agile and dialectical manner, ensuring that the process is cognitively light for the human coach, and that the knowledge acquired by the machine is – by design and provably, in a formal sense – acceptable to the human with whom the machine is interacting.

References

1 Loizos Michael. Machine Coaching. In: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI), pages 80–86, Macao SAR, P.R. China, August 2019.

3.10 Social biases: Identifying Stereotypes about Women and Immigrants – The Case of Misogyny

Paolo Rosso (Technical University of Valencia, ES)

Language has the power to reinforce stereotypes and project social biases onto others. At the core of the challenge is that it is rarely what is stated explicitly, but rather the implied meanings, that frame people's judgments about others. This is the case of stereotypes about women and immigrants, two social categories that are among the most preferred targets of hate speech and discrimination. In the first part of the talk, we address the problem of automatic detection and categorization of misogynous language in social media. Special emphasis is given to the issue of transparency during data collection and labelling, as well as at the time of the explanation of the categorization of misogynous language. Finally, we illustrate a taxonomy that we proposed to address the problem of stereotypes about immigrants.

References

1 Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In 13th International Workshop on Semantic Evaluation (pp. 54-63). Association for Computational Linguistics.

2 Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian, 12, 59.

3 Sánchez-Junquera J., Chulvi B., Rosso P., Ponzetto S. (2021) How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. In: Applied Science, 11(8), 3610 https://doi.org/10.3390/app11083610

4 Working groups

4.1 Stakeholders and their Questions (day 1)

Judy Kay (The University of Sydney, AU), Tsvi Kuflik (Haifa University, IL), and Michael Rovatsos (University of Edinburgh, GB)

The online nature of the seminar necessitated a setup where sub-groups in compatible timezones convened for extended sessions twice each day, with those meeting in a larger session then further collaborating in small breakout groups on specific topics. The participants of each session were re-shuffled every day to maximise opportunities for all of them to interact with each other at least for part of the event. Further, to allow all participants to engage with the material presented, recordings of invited talks presented in each session were provided for the benefit of those not able to participate in those sessions due to their geographical location. The working groups across all sessions also collaborated asynchronously through shared online documents, documenting discussions and enabling cross-fertilisation between the work of individual groups.

This process was followed for the first three days of the seminar, with the final day being structured around working groups forming around topics for different sections of the joint "transprency by design" framework document authored jointly by all seminar attendees.

On the first day, working groups focused on a user-centric perspective to understand requirements for transparency. For this purpose, participants were split into small groups that aimed at identifying the questions users want to be answered when considering the use of algorithmic systems.

At a high level, these questions characterise the sorts of questions that a transparency by design approach will enable key stakeholders to answer. The group discussion was organised in six stages:

Stage 1 Select one key context that the group has some expertise and interest in, and identify the key classes of stakeholders.

Stage 2 For one such stakeholder group, we brainstorm to define an initial broad set of driving questions the people in that stakeholder group may want to be able to answer.

Stage 3 Repeat the above steps for a second stakeholder group.

Stage 4 Identify the questions that the group considers to be most important to support in the design process, taking account of additional stakeholders identified in Stage 1 – this was the narrowing phase of the brainstorm discussion.

Stage 5 Review the conclusions across the other groups of participants (reviewing the shared document used by all groups). Each group then used this to both refine the set of driving questions for that group's context and to identify the similarities and differences between them.

Stage 6 Finalize what to share with other groups.

The groups then returned from their breakout groups to a joint session, with a representative from each group presenting the key outcomes of their group's work. There were in total six groups and they selected three different domains, in the areas of education (3), job recruitment (2) and e-commerce (1).

For *job recruitment*, the identified key stakeholders are government regulators, certification agencies, users (job consultants, applicant (not) offered a position, lawyers contesting decisions, software developers, analysts, and hiring entities). The most important stakeholders interacting with a job recruitment system that were identified were regulators and applicants.

The key questions concerning the transparency of the system were "how", "why" and "what". There were two major points of view, applicant and job consultant, but the questions that were identified are relevant to both. For the "how", these were: How is the candidate being scored? How is the user's privacy protected? How much control do the users (employers/candidates) have to appeal/contest the decision? How well can the system handle discrimination bias? How can the candidate be a good fit for a specific job position? How does the system make a decision? For the "why", they were: Why did the candidate not get the job position while someone else was successful? For the "what", they were: What is the appropriate process followed for developing the software? What control do the end users have over the system (e.g. once we have identified that the system is doing something wrong, is the candidate able to change this)? What criteria were set for the system to operate? What concepts, values, terms have been defined in the system? What data was used to train the model? What process was used to make the decision to use that data? What are known historical good and bad practices? What went wrong in the past and why (no matter whether/what technology was used)? Have we investigated this history? What have we learned from these investigations?

For *education*, the key stakeholders are the students, teachers, professors, the parents, the university and school administrators, IT experts (technology developers and operators of the assessment system) at the institution, research funders and society at large (general public, journalists, civil society activists). The most important stakeholders identified in an educational system are the teachers and students. Again, the key questions were divided into "why", "how" and "what". The key questions concerning the transparency of the system are "why" questions, such as: Why does the system make an intervention (which can be justified based on the system's objective), or why is this problem (specific topic/concept) recommended for the student to solve? In terms of "what": What criteria does the system use to make a decision (e.g. Why did I get this grade? What are the marking criteria?)? For the "how", the questions are about a variety of aspects regarding process and outcomes: How did the system select a peer group to compare the student's performance to? How is a specific intervention generated? How would the grade differ compared to human-given grades? How can I improve my grade? How does the system produce the specific grade or decide the student's level? How does an adaptive educational system determine/evaluate student knowledge for a domain topic or concept? How can we ensure consistency of assessment? How do I know that the accuracy of the assessment is correct? For the "what" the focus was on: What information is being shared with third parties (external vendors)?

In the *e-commerce* domain, the key stakeholders are buyers, sellers, creators – i.e. music track, performer vs. producer, advertisers, developers of a system, shipping/supply chain, recipients, regulatory (depending on product) operator. Here the main stakeholders are buyers and sellers and as they greatly differ, as do the questions. The key questions regarding the transparency of the system that will concern buyers are why they get (or do not get) specific recommendations? How to interface with the system, and how the system deals with any ethical and privacy concerns? How reliable/trustable the information and reviews of a product are that are presented? How people build their mental models of the space of products and services? The driving questions of the sellers concern the transparency of why their product is or is not recommended to specific buyers, how the pricing works, how the advertisement of products is performed, what responsibility issues there are; and what control a buyer or seller interacting with the system has. In general, the most important question is why someone gets a specific recommendation, what control buyers/sellers have while interacting with the system, and how trust in the system has been calibrated.

Through discussions in the working groups on the first day, a comprehensive account of requirements for transparency was elaborated for a number of different scenarios (including additional ones beyond those described above). These demonstrated very clearly that the breadth of possible concerns is potentially too great to condense into concise models, especially considering the specifics of different domains, and the societal, regulatory, legal, and ethical parameters that depend on the context of use of algorithmic systems in each sector. Nonetheless, it became clear that there are ways to systematically analyse who needs to know what about a system, when and how this information should be delivered to them, but also that these are issues that are not currently routinely addressed when algorithmic systems with potentially wide-ranging societal impacts are developed.

4.2 State of the Art and Emerging Priorities (day 2)

Judy Kay (The University of Sydney, AU), Tsvi Kuflik (Haifa University, IL), and Michael Rovatsos (University of Edinburgh, GB)

A major outcome of discussions on the first day of the seminar was the need to focus on collecting and analysing data in order to answer transparency-related questions. Framed within an exploration of the state of the art and research challenges around making progress in this direction, the second day followed a discussion format where working groups progressed through four stages: 1) Describing the state of the art on data collection in each of the domains discussed; 2) identifying gaps in terms of data collection; 3) considering requirements for an agile test-driven environment to bridge these gaps; and 4) identifying challenges that need to be addressed to enable such methodologies. The domains discussed in the six groups were similar to those of the previous day: e-commerce, job recruitment, and educational systems.

In the e-commerce domain, the main questions from the previous day were: Why did a customer get this specific recommendation, what control does the customer have over who sees the seller's product, and why are features to allow such customer-side control configured the way they are. Transparency in recommender systems is usually handled via providing explanations for recommendations in current systems, but many publications are also emerging that address fairness in recommendations. Looking at Amazon as a prime example, the user model can be built given based on short-term profile elements: current session interaction (items clicked, time spent on reading reviews/exploring a specific item, sequence of items), which is often combined with long-term data (the previous history of the user including user and item contexts, the behaviour of similar users – collaborative filtering – and derived attributes). Some of this data involves private self-reported information and some involves public data. Explicit data can also be collected from user ratings or surveys. Such data collection leads to additional questions regarding data transparency, e.g. how the specific user model is built and how it is used later on for recommendations, what training data is used, from what time period, whether different algorithms are used in different countries by the same company due to different legislation in those countries, etc. Some data pre-processing methods that can be used on the collected data such as adversarial inputs, detection of skew in available data for popular vs. less-popular items, data augmentation techniques and pre-trained embeddings for text or image analysis. The gaps from the existing literature regarding the data pre-processing techniques are that transparency issues of the data still exist especially when using pre-trained embeddings for text or image processing as well as a lack of transparency from technology providers regarding underspecification of how their systems operate, including how they may have been "sanity checked" and "stress tested".

Taking all these into consideration, and using an agile test-driven approach, one of the two groups suggested an approach where, at each software development cycle (requirements, design, implementation and testing) various stakeholders and technologies, appropriate for that specific stage, should be used. For example, for the specifications/requirements analysis phase the main stakeholders are end users, regulators and owners, and technologies/methods used are standard software engineering techniques. However, for a future transparency-driven development, where users would like to get both an explanation about the model itself and its outcomes, and to be able to see what features were used to build her actual user model, the following approaches should be considered: In the requirements stage, transparency could be defined as a non-functional requirement that might later become a functional requirement (for example, how to present the user profile/model to the user with features that were used for building it, e.g. via user preferences, with a checklist of relevant features the user can opt out of). This should be in addition to providing model explanations, as well as outcome (recommendation) explanations and global explanations for the overall software quality assessment.

In the testing stage, the generation of test cases to validate transparency requirements will be important (e.g. in regression testing), where the testing approach itself should also be transparent. In terms of stakeholders, it should be clear who is involved in each stage of the engineering process, what decisions they have to make, who is responsible for each decision, and whether these decisions need to be reported to regulators. Regarding the technologies used, those used during the certification processes should be independent of those used in development.

The group highlighted the importance of educating stakeholders in order to embed such an approach in systems development, but some open questions remain unresolved, including: What certifications are available in terms of trusting the system? Does the public want to be educated to understand such certifications? From an ethical/legal perspective, users need to know their rights. Were the engineers trained in transparent systems design, and how should this issue be addressed (e.g. should such training be legally mandated)?

Another group performed a literature survey on state-of-the-art research related to transparency in data collection in the area of recruitment. They identified papers and systems that deal with automatic recruitment and its implications, challenges, and solutions. Most of the papers referred to fairness in the recruitment systems, which in some sense overlaps with transparency. However, it is hard to identify applications implemented in the real world. In addition to this, some solutions proposed in the literature are not closely embedded in the context of hiring and recruitment, and are often more about generic machine learning based systems. For systems that deal with people and decisions taken about them, the questions seem to be similar. Moreover, there are generic solutions that exist and maybe we need to look into them and one should consider whether there is a need for adaptation towards objectives we might aim for in the recruitment domain.

In the context of educational systems, the questions for transparency concern mostly what kind of data the system has about students, how it ranks them, why a student obtained a particular recommendation, and how accurate this recommendation is (which is typically also asked by the educational provider). The group found many publications that include use cases such as machine learning systems that predict learning outcomes and forecast student grades. Another aspect discussed were privacy concerns around student data, as evident from two case studies (one on how Coursera handles GDPR issues, and another one using Named Entity Recognition to anonymize student data). In these case studies, the data used as input to algorithmic components includes demographic data, student grades, their responses in assignments, evaluations of teaching performance, represented either as text or in tabular data. Techniques that are usually used to pre-process the data included data cleaning techniques for the training data that would be used in a classification or clustering algorithm, web mining techniques, the use of association rules to mine learner profiles and diagnose learners' common misconceptions, sequential pattern mining to extract and present patterns that characterize the behavior of successful groups, and web mining techniques that group documents according to their topics and similarities and provide summaries. As for transparency, a case study that reports on using the LIME explainability method for student monitoring techniques used in a recommender system, and the QII approach for providing transparency reports in learning systems stood out as more principled studies in the state of the art.

Regarding the allocation of people to positions, job applications discussion focused on problems of both the regulators and the applicants, and, in particular, on helping the applicants to understand why they did not get a job. This is not only about looking for discrimination, but also about coming to a greater sense of self knowledge and job knowledge and and bringing all these pieces together. Since both regulators and naive users are involved in this process, an approach similar to digital forensics seems appropriate, not only for the regulators to ensure fair treatment, but also for the applicants who should be able to inspect differences between themselves and other, successful, applicants. Another important concern is personal data retention by platform providers, in particular how long personal data can be kept by the company and used. This is important not just because of GDPR, but also because of labor markets and social policy. Apart from understanding how an applicant is different from others, it is also essential to be able to identify whether there *are* appropriate jobs they can apply for with high chances of success, but also to help them understand how they can increase their chances. In other words, transparency should also be used to empower users to make their own appropriate choices rather than just explaining current systems behaviour.

Finally, in the college admissions context, further dimensions where discussed included explaining whether diversity targets in the admissions process are met and outcomes improving, understanding which students need/deserve more detailed explanations, and the wider issue of understanding which stakeholders require what different types of system output, including predictions and explanations. Concerning the data collection process, the system might use self-reported demographic and academic performance data that candidates enter in their applications, but input data might also include information from social media and the candidate's other online activities. Data pre-processing includes data cleaning and "munging", and also the use of feature engineering methods. In terms of data management, in state-of-the-art approaches, data provenance techniques are usually employed, and privacy policies may apply differently to different datasets and data sources. Open questions in this domain include: What are the important user characteristics and how are those measured, especially if they go beyond academic performance? Why is a particular piece of data being used and what is the evidence for it being relevant to the successful completion of the degree? The remaining gaps from the state-of-the-art concern the acquisition of all the necessary data to be in an accessible and usable format, and methods to assess the adequacy and validity of the system.

To summarise, from all the contexts chosen by the groups, there is a wide range of state-of-the-art publications on transparency in the respective domains. Concerning the data collection process, in almost all recommender systems, both implicit and explicit information are collected from users. Transparency is achieved using explanations for provided recommendations. In educational systems, the data collection process depends on the objective of the system but usually it concerns self-reported data (e.g. demographic data, responses to assessment questions) or information given by other users (e.g. student grades). Data pre-processing techniques might include the general data cleaning techniques or more specific techniques such as testing against adversarial inputs, web mining and data augmentation techniques. Also, various stakeholders of the system should be considered and explanations provided to them should differ to meet the particular needs of each stakeholder. In terms of implications for software engineering practice, the case studies demonstrate that transparency should be considered at the requirements phase and also throughout the whole software development lifecycle, which is not the norm in current practice. In particular, the shift to data-driven system mandates more rigorous planning, monitoring, and documentation of data provenance, collection, and analysis methods deployed.

4.3 Thematic Research Challenges (day 3)

Judy Kay (The University of Sydney, AU), Tsvi Kuflik (Haifa University, IL), and Michael Rovatsos (University of Edinburgh, GB)

Following on from working groups convened on day two, the third and final day of structured group work focused on mapping the existing research landscape to further identify gaps and seek to articulate an agenda for future research themes; with the aim of pulling results from all three days together to feed into a jointly authored programmatic paper that would provide the foundation for a "Transparency by Design" methodology in the subsequent final seminar day.

As before, working groups were asked to choose a concrete use case context to focus on, in order to identify those questions that are most pertinent to the chosen context, and explore state-of-the-art design and implementation methods which then lead to the identification of existing gaps. A specific focus for this day were challenges around building interfaces that can support users in scrutinising AI algorithms and pinpointing the challenges surrounding the creation of such interfaces. The discussions undertaken by the six groups covered some more general concerns on recommender systems as well as specific issues from one of the following domains: applicant ranking in an educational and employment context, e-commerce systems, care robots, and debugging tools for data-driven/AI systems, which are a cross-cutting concern for many domains.

Regarding systems that rank applicants and items, the following seem to be lacking: (a) an ability to trace the reasoning of deep learning systems, (b) the ability to explain to end users what features of their profiles were used for ranking, (c) the ability to opt in and out of the usage of certain features by the systems, (d) suggestions on what to do for lower-ranked users, and, as a more challenging gap, (e) the development of intrinsically transparent ranking methods, where transparency is not just provided *post hoc*. In the context of interfaces and explanations offered for rankings, common everyday life applications in social media

(Facebook), e-commerce (Amazon), video streaming (Netflix) and recruitment (LinkedIn) applications were discussed. In the example of video streaming platforms, services only show similar movies and percentages without explaining them. In terms of gaps, one major issue seems that it is impossible to find examples of interfaces for stakeholders other than the end users. Across many platforms, explanations seem to be very short and basic, and are likely to be insufficient for the information needs of at least some users, let alone other stakeholders such as analysts, regulators, and policy makers; where more detailed explanations also offer the opportunity to be open to end users as a side-benefit.

In the area of *e-commerce*, two main questions were identified: *Why am I getting a particular recommendation (consumer-side)?* and *Who sees my product (seller-side)?*. Explanation types vary across different types of data and the same explanation can be offered in different ways depending on the target audience. In this area, the core function of the system is to produce recommendations, so one way to look at the question is what type of explanations to use for this purpose. The areas of identified gaps included (1) systematic approaches and methodologies for designing transparency-compliant systems (e.g. training courses, regulations and other resources from relevant authorities, personalised explanations), (2) privacy and stereotyping related issues (e.g. understanding the trade-off between privacy and personalised explanations), (3) auditing and certification (e.g. regular and ongoing auditing). When it comes to user interfaces there is a lack of multimodal explanations and discoverability of existing explanations. From an ethical point of view, it is important to ask how not to abuse explanations and the interface to manipulate people (or for the platform to be manipulated by its users).

In the field of *care robots*, the main questions identified were: What is the aim of building such technology, who is meant to be the beneficiary, who are the explanations for, how are decisions taken by the robot (which can change the relationship between the carer and the patient) and what level of transparency is needed during the development, data collection, certification phase and deployment of such robots? Multiple potential areas of transparency were discussed, such as the transparency of the recommendations provided by the robot (i.e. its reasoning behind concrete decisions), transparency of how the system exactly operates and transparency for regulatory purposes. It was suggested that different levels of transparency are needed for different areas but participants noted that in certain cases transparency can lead to privacy concerns. For instance, capturing the original design and development conversations can be useful in order to demonstrate due diligence but also can endanger privacy and cybersecurity of the users of the robots. The capturing of abstracted but comprehensible levels of development and operations information is necessary for the maintenance, inspection, installation and explaining of the system. Since care robot systems give advice on high-stake matters, it is very important that the advice can be traced back to the development process as well as the medical literature and research underpinning their design.

The discussions in the working groups on this day revealed that the landscape of methods proposed by the academic literature on various sub-topics related to transparency is indeed very rich, and ranges from new advances to improve the explainability properties of AI algorithms and fairness-aware recommender system design all the way to work on advanced user interfaces that enhance the understandability and configurability of algorithmic systems. However, the current literature not only presents itself as fragmented in terms of providing complete accounts of end-to-end transparency-driven approaches to systems development, there is also a (much more serious) lack of connecting the specific research advances to actual systems in the real world, where very little is known about how they are actually implemented. Given this, it is unsurprising that some of the existing work seems hard to apply to real-world examples, with authors making often unrealistic assumptions that can only be fulfilled in experimental systems developed purely for research purposes. Overall, it seems that, while some technical challenges remain, the methodological building blocks for developing a holistic transparency by design approach are largely in place. What is missing is an ability to map them onto a context-specific methodology that could be embedded in the development of impactful (usually commercial) real-world systems. In this regard, the most important roadblock as the moment is an inability to evaluate these systems to propose concrete transparency improvements. There are significant barriers to overcome here in terms of cross-sector collaboration and education, and our overall longer-term objective of proposing a methodological framework might help accelerate the dialogue between researchers, developers, regulators, and users in this respect.

4.4 Toward a Transparency by Design Framework (day 4)

Judy Kay (The University of Sydney, AU), Tsvi Kuflik (Haifa University, IL), and Michael Rovatsos (University of Edinburgh, GB)

During the fourth and final day, the results of the first three days were discussed and summarised into a joint document that is intended to form a basis for a joint paper. The structure of the document followed the results of the discussion of the topics and the order of the discussion in the first three days, organised into the following chapters:

1. Why transparency?
2. Challenges posed by current algorithmic systems
3. Transparency-enhancing technologies
4. Transparency by design – principles
5. Transparency by design – methodology
6. The Road Ahead

Throughout the day, and also after the conclusion of the seminar, self-selecting sub-groups of seminar participants collaboratively developed material for each individual section, working synchronously and asynchronously. The work of these groups was guided by an overarching structure and guidelines for for the content of the planned paper, following a set of key questions that guided the discussion, which we list below together with key results of the discussions:

- Key questions to answer about “Why Transparency”:
  1. What is transparency and why is it important?
  2. How does it relate to other concerns around AI?
  3. Why is it timely to develop new transparency capabilities?

The high-level results of this discussion included that there are many public policy documents that call for transparency, but that different experts take very different perspectives and talk about different things when it comes to transparency. Nonetheless, there is some universal agreement on the importance of transparency for trust and its calibration, in order to have accurate mental models of systems for users, to support accountability and auditing, and to enhance fairness and human control. It was noted that transparency should not be taken as a panacea, or an end in itself. Rather, transparency can contribute to ensuring human accountability, whether this comes from the use of good design and development practices, addressing –or even readdressing– problem issues, or to underpin appropriate levels of legal liability.

- Key questions to answer about "Challenges posed by current algorithmic systems":
  1. What has changed in current systems that creates challenges for transparency?
  2. How hard will it be to enable pervasive transparency given these challenges?
  3. Why does this matter to users, organisations, societies?

The views discussed here highlighted that the key changes that create new challenges are due to a growth in the ubiquity of AI-based systems, the presence of particular impacts of these system on vulnerable groups and the growing cost these incur, but also the increasing availability of tools for creating these systems, which accelerate these impacts. This can be seen as a consequence of the "democratization" of AI, which makes it so readily available that it is incorporated in many systems. Furthermore, data acquisition has deeply changed over time, from a time when the AI system did not know anything that was not inserted in it by a human directly to modern-day machine learning using global-scale Internet data, which makes it harder to audit for and filter content, with the real risk of poor data and associated inferences. A number of increasingly prevalent practices can give rise to further transparency risks: The use and reuse, including sometime on a large scale, of datasets that lack proper documentation; the development of "general" pre-trained models as "cognitive services", which developers can use in a plug-and-play manner; and the emergence of application-agnostic components and tools released in the public domain without knowing or being able to control their downstream usage. On a societal level, there is a growing demand for accountability at the level that would be expected of people making the decisions that are entrusted to systems, and this is reflected in emerging legal and regulatory requirements, together with a growing recognition of needs and challenges of education about AI, including AI literacy of the general public and curricula for schools. Participants identified many factors that make it challenging for developers to support transparency, including: The prevailing attitudes of those responsible for major players; the technical difficulty of even defining what we mean by transparency, much less enabling it; challenges related to designing for multiple stakeholders; the potential for adversarial interaction to game systems that might arise from increased transparency; and a lack of understanding of the complex dynamics of human-AI systems. This presents us with a formidable challenge, because of the many costs of the current lack of transparency, which may entrench disadvantage and inequality through a broad range of unanticipated societal impacts. It was noted that it is extremely hard to find general solutions to transparency problems. Domain-specific analysis is key to be able to do solid assessments of potential solutions specific to the use case. This is particularly important as, more often than not, the multiple stakeholders have different conflicting objectives.

- Key questions to answer about "Transparency-enhancing technologies":
  1. What approaches to enhancing transparency have been proposed?
  2. What benefits did they deliver, and what were their limitations?
  3. What are the transparency issues they did not address?

In the discussion of these questions, participants focused on several issues. The first one is considering all stages of the design and development process of AI-based systems, including data preparation, the learning algorithm/model/process used, model and outcome explanations, and user interfaces through which these will provided. Along a second dimension, each of this issues needs to be considered for all relevant stakeholder groups, from the model and software developers to project managers and end users. This needs to be done in a fine-grained way, as, for example, different types of end users may have very different requirements. The participants discussed a range of methods available (and missing) from the literature, highlighting the importance of distinguishing between different types of data (e.g. text, images, video, etc), its source (e.g. offline data collection vs. crowdsourcing), the annotation processes used, traceability and provenance methods used, fairness metrics applied. They discussed differences between types of learning algorithms used, e.g. supervised, semi-supervised, unsupervised, and reinforcement learning methods, and to what extent different explainability methods can be applied to them. For the latter, there already exist useful taxonomies of explainability methods (including rule-based white-box models, surrogate models of black-box models, and counterfactual explanations), and there is evidence that users favour "why" and "how" explanations along general lines, but that one-size-fits-all explanations are not always effective or desirable. Finally, in terms of interfaces, there is a body of work on these especially in the area of recommender systems, which include visualisation of latent influencing data and variables, the explication of intermediate results, but also new interfaces that allow improved user control and scrutability.

- Key questions to answer about "Transparency by design – principles":
  1. What are the high-level guiding principles we want to embed in Transparency by Design?
  2. What is the justification behind them? (Why do we think they are important?)
  3. How realistic is it that they could be achieved? (Is this aspirational or practical?)

Under this theme, the participants broke down the overall question into issues relating to defining transparency by design in terms of its scope, purpose, and method. The first question here is to consider where transparency should be enabled, focusing on the design process, which should explicitly state assumptions and context, trace the provenance of data used and how it flows through a system, make the construction process of models and the use of their outcomes explicit, and document testing and validation processes and outcomes. The second is clarifying the purpose of transparency, which should be to empower users (enabling contestability, data subject rights enforement, configurability of systems, and their risk management). Transparency by design approaches will need to effectively demonstrate they address the requirements of different stakeholders in this regard, including customers, regulators, developers, system integrators. In term of the methods underpinning transparency, a framework will have to specify how they are evaluated and validated (e.g. through benchmark tasks and metrics), establish principles for ensuring that the relevant bodies and individuals will and can understand the information provided, and provide a normative framework for understanding when explanations are appropriate, who is responsible for providing it, and what the appropriate ways of communicating the information are.

- Key questions to answer about "Transparency by design – methodology":
  1. What changes need to be made in each stage to satisfy transparency principles?
  2. Who are the people that need to take responsibility for applying these techniques?
  3. Can we describe these techniques through a relatively simple process model?

The high level results of the discussion in this area reiterated the importance of taking a holistic approach, where transparency would apply to input data, learning process, decision making, and output. A solid methodology would include a full account of decisions made by the system (or by humans using it) through a range of explanation methods using transparency-enhancing user interfaces. Participants proposed a preliminary "five star" framework of transparency, building on the following pillars: (1) explanation of decisions; (2) stakeholder-appropriate provision of these; (3) user interfaces that support transparency and exploration of the system; (4) open sharing of validation process and results; and (5) comprehensive description of the input data.

- Key questions to answer about "The road ahead":
  1. What are key problems the research community needs to focus on?
  2. What kinds of experts (academic/industry/government) are needed to do this work?
  3. What other enabling steps need to happen, e.g. in industry or government?
  4. What are the limits of TBD, what problems will it not solve?

For this final them, the groups discussed and contributed insights on suitable ways to further pursue a Transparency by Design agenda. One important aspect that was raised here was the importance of advancing general principles for corporate, legal, and product transparency in AI. This will require enhancing the wider understanding of existing regulatory obligations, outreach and education to legislators, enforcers, civil society watchdogs, judges, developers and tech corporations. Funding and political will to enforce societal expectation will be important, together with substantial efforts to counter misinformation emanating from AI-based systems, which is recognised as one of the major problem areas. It is important to acknowledge existing initiatives, including proposals for regulatory frameworks and standards. Undoubtedly, a lot of further development work and research will be necessary, which might even lead to new types of professional roles for people supporting AI transparency with responsibilities for oversight, monitoring, and enforcement, embedded in different types of organisations. Before effective transparency capabilities are developed across the industry, we will likely see much more work on standards, exchange of good practice, the resolution of tensions between transparency and intellectual property rights, and, of course, technical development of new transparency-enhanding technologies and interfaces.

Participants

- Elisabeth André
Universität Augsburg, DE
- Bettina Berendt
TU Berlin, DE
- Nehal Bhuta
University of Edinburgh, GB
- Maria Bielikova
KInIT – Bratislava, SK
- Veronika Bogina
University of Haifa, IL
- Joanna J. Bryson
Hertie School of Governance – Berlin, DE
- Robin Burke
University of Colorado – Boulder, US
- Aylin Caliskan
George Washington University – Washington, DC, US
- Carlos Castillo
UPF – Barcelona, ES
- Ewa Cepil
UPJPII – Krakow, PL
- Cristina Conati
University of British Columbia – Vancouver, CA
- Aviva de Groot
Tilburg University, NL
- Gianluca Demartini
The University of Queensland – Brisbane, AU
- Virginia Dignum
University of Umeå, SE
- Fausto Giunchiglia
University of Trento, IT
- Riccardo Guidotti
University of Pisa, IT
- Meeri Haataja
Saidot Ltd – Espoo, FI
- Judy Kay
The University of Sydney, AU
- Styliani Kleanthous
Open University of Cyprus – Nicosia, CY
- Ansgar Koene
EY Global – London, GB
- Joshua A. Kroll
Naval Postgraduate School – Monterey, US
- Antonio Krüger
DFKI – Saarbrücken, DE
- Tsvi Kuflik
Haifa University, IL
- Bob Kummerfeld
The University of Sydney, AU
- Loizos Michael
Open University of Cyprus – Nicosia, CY
- Antonija Mitrovic
University of Canterbury – Christchurch, NZ
- Kalia Orphanou
Open University of Cyprus – Nicosia, CY
- Jahna Otterbacher
Open University of Cyprus – Nicosia, CY
- Anna Perini
CIT- FBK – Povo, IT
- Lena Podoletz
University of Edinburgh, GB
- Iris Reinhartz-Berger
University of Haifa, IL
- Paolo Rosso
Technical University of Valencia, ES
- Michael Rovatsos
University of Edinburgh, GB
- Avital Shulner-Tal
University of Haifa, IL
- Alison Smith-Renner
University of Maryland – College Park, US
- Andreas Theodorou
University of Umeå, SE
- Vincent Wade
Trinity College Dublin, IE
- Emine Yilmaz
University College London, GB