Report from Dagstuhl Perpectives Workshop 22262

Human-Centered Artificial Intelligence

Wendy E. Mackay[*1], **John Shawe-Taylor**[*2], **and Frank van Harmelen**[*3]

1 **INRIA Saclay - Orsay, FR.** wendy.mackay@inria.fr
2 **University College London, GB.** jst@cs.ucl.ac.uk
3 **VU University Amsterdam, NL.** frank.van.harmelen@vu.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Perspectives Workshop 22262 "Human-Centered Artificial Intelligence".

The goal of this Dagstuhl Perspectives Workshops is to provide the scientific and technological foundations for designing and deploying hybrid human-centered AI systems that work in partnership with human beings and that enhance human capabilities rather than replace human intelligence. Fundamentally new solutions are needed for core research problems in AI and human-computer interaction (HCI), especially to help people understand actions recommended or performed by AI systems and to facilitate meaningful interaction between humans and AI systems. Specific challenges include: learning complex world models; building effective and explainable machine learning systems; developing human-controllable intelligent systems; adapting AI systems to dynamic, open-ended real-world environments (in particular robots and autonomous systems); achieving in-depth understanding of humans and complex social contexts; and enabling self-reflection within AI systems.

Seminar June 26–July 1, 2022 – http://www.dagstuhl.de/22262

2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI); Human-centered computing → Interaction design; Mathematics of computing → Probability and statistics; Theory of computation → Probabilistic computation; Theory of computation → Automated reasoning; Theory of computation → Constraint and logic programming; Theory of computation → Machine learning theory; Theory of computation → Algorithmic game theory and mechanism design; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning

Keywords and phrases Human-centered Artificial Intelligence, Human-Computer Interaction, Hybrid Intelligence

Digital Object Identifier 10.4230/DagRep.12.6.112

1 Executive Summary

Wendy E. Mackay
John Shawe-Taylor
Frank van Harmelen

This workshop brought together 22 participants with a diverse background from AI, Robotics, HCI, Ubiquitous Computing, Business and Sociology, from across Europe and North America laid the groundwork for a manifesto on Hybrid Human-centered AI systems.

Informed by currently ongoing large initiaves such as the EU-funded Humane AI Net, the Dutch Hybrid Intelligence Center, the Danish Centre for Hybrid Intelligence, and OECD AI policy framework, four pillars of the manifesto emerged: (a) Collaboration and Cooperation,

[*] Editor / Organizer

Human-Centered Artificial Intelligence, *Dagstuhl Reports*, Vol. 12, Issue 6, pp. 112–117
Editors: Wendy E. Mackay, John Shawe-Taylor, and Frank van Harmelen
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

(b) Control & Adaptivity, (c) Transparency & Explainability, and (d) Societal dimensions. For each of these pillars, the workshop resulted in (i) key terminology, (ii) key research questions, (iii) metrics and methodologies, and (iv) benchmarks and challenges.

The above resulted in a solid framework for the Hybrid Human-Centered AI manifesto to be written by the partcipants in the months following the Dagstuhl workshop.

2 Table of Contents

3 Overview of Talks

Wendy Mackay presented the Human-Computer Interaction approach to taking a human perspective when interacting with an intelligent system. She explained that the focus is not on the human alone, or the system alone, but rather the interaction between them, where interaction is treated as a phenomenon in its own right that can be designed, studied and controlled. She explained the cycle in the natural and social sciences between theoretical and empirical research, and how to incorporate human-designed interactive systems into the research process. She briefly described generative theory for creating novel interactive systems, as well as the possible relationships between users and interactive systems, and how they blur when designing human-centered AI systems. Finally, she described a number of human-centered participatory design methods, and the importance of prototyping how humans will interact with a proposed intelligent system, including explicitly identifying, illustrating and reflecting on possible breakdowns.

Marko Grobelnik presented the OECD AI policy framework. He discussed issues of human vs. AI autonomy, and questioned whether the OECD framework is too reductionist, placing the human within a machine framework, rather than the other way around. He also discussed the AI system lifecycle and AI system classification.

Frank van Harmelen presented the research agenda for the Dutch Hybrid Intelligence Center, with 80 Ph.D. students over 10 years (slides here). He explained the project's CARE goals: Collaborative (synergy with humans), Adaptive (adapt to humans and the environment), Responsible (ethical performance), and Explainable (share and explain awareness, goals and strategies).

Paul Lukowicz presented the HumanE-AI European research network.

Paul Lukowicz presented two overall frameworks. The first framework includes: Communication; Oversight/Control; and Frame of reference. The second organizes the Outcome space in terms of physical effect, impact on humans, and impact on society.

4 Plenary discussions

After the opening sessions on Monday morning, each participant gave a three-minute presentation of their background, goals for the workshop, and key issues they would like to address in the workshop. The group is highly diverse, with representatives from AI, Robotics, HCI, Ubiquitous Computing, Business, Sociology, from across Europe and North America.

In the Monday afternoon session, the group identified six major issues to address, which were compressed into four topics for breakout groups: Collaboration and Cooperation, Control & Adaptivity, Transparency & Explainability, and Societal dimensions, to be discussed in greater detail in four breakout sessions during the week. Each breakout session was encouraged to discuss the following topics for the manifesto: Key Terminology; Research Questions; Methods, Competences and Frameworks; Benchmarks, Moonshots and Challenges; and Policy.

5 Working groups

In working groups on Monday afternoon, the participants addressed the following assignment:

1. Choose a target user and a context where a user or (users) interact with an intelligent system, based on one of the four morning session topics;
2. Develop a scenario and describe prototypes (on paper, or otherwise) to illustrate the interaction;
3. Shoot a video prototype of a realistic scenario that illustrates how the user(s) interact with the system, including at least one situation where the system breaks down.

Each of the five groups then presented and discussed their video and the challenge presented by the breakdown.

In working groups on Tuesday afternoon, participants discussed the issues related to Collaboration and Communication in Human-centered AI. At the end of the afternoon, participants met in a plenary session to discuss the results of each breakout session.

In working groups on Wednesday morning, participants formed four breakout groups with specific examples related to the topic of Transparency and Explainability. At the end of the morning, participants met in a plenary session to discuss the results of each breakout session.

In working groups on Thursday morning, participants formed four breakout groups to discuss specific examples with respect to the dimensions outlined in Paul Lukowicz' presentation (mentioned above). At the end of the morning, participants met in a plenary session to discuss the results of each breakout session.

In working groups on Thursday afternoon, participants formed three breakout groups, one focused on policy issues, the remaining three groups worked on writing the following:

1. Definition of HCAI,
2. List of provocative statements for the manifesto,
3. Identification of leading examples
4. Suggested Ph.D. topics

In a plenary session on Friday morning, the group discussed creating four concrete examples that illustrate the design space of human-centric AI and settled on four each with an assigned author: Interactive collaborative music, Social Media, Decision support for Health, and Crisis Management.

After organising the collected breakout notes into the key components of the manifesto (terminology, research questions, methods and metrics, benchmarks and moonshots, and policy), the participants then broke up into working groups and drafted the key sections of the manifesto

In a final plenary session the group then discussed the results and future directions, including transforming the manifesto into a 10-20 page journal article, with scientists as the target audience, contributing to the HHAI conference in Munich next year, and establishing a Human-centered AI Master's program.

Participants

- Michel Beaudouin-Lafon
University Paris-Saclay – Orsay, FR
- Mehul Bhatt
University of Örebro, SE
- Stefan Buijsman
TU Delft, NL
- Mohamed Chetouani
Sorbonne University – Paris, FR
- Ulises Cortés
UPC Barcelona Tech, ES
- Adam Dahlgren Lindström
University of Umeå, SE
- Emmanuelle Dietz
Airbus – Hamburg, DE
- Marko Grobelnik
Jozef Stefan Institute – Ljubljana, SI
- Alípio Jorge
University of Porto & INESC TEC – Porto, PT
- Antonis C. Kakas
University of Cyprus – Nicosia, CY
- Samuel Kaski
Aalto University, FI
- Janin Koch
INRIA Saclay – Orsay, FR
- Helena Lindgren
University of Umeå, SE
- Paul Lukowicz
DFKI – Kaiserslautern, DE
- Wendy E. Mackay
INRIA Saclay – Orsay, FR
- Ozlem Ozmen Garibay
University of Central Florida – Orlando, US
- Janet Rafner
Aarhus University, DK
- Laura Sartori
University of Bologna, IT
- John Shawe-Taylor
University College London, GB
- Jacob Sherson
Aarhus University, DK
- Marija Slavkovik
University of Bergen, NO
- Philipp Slusallek
DFKI – Saarbrücken, DE
- Frank van Harmelen
VU University Amsterdam, NL
- Katharina A. Zweig
TU Kaiserslautern, DE