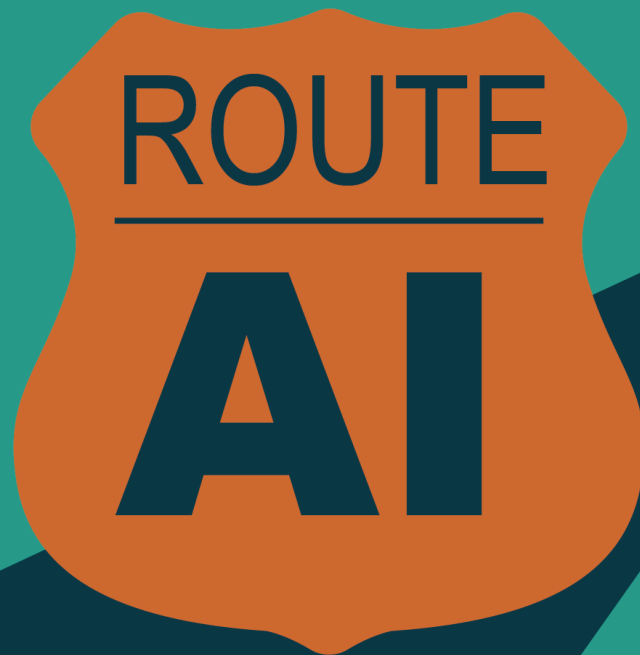


# Towards *Reflective* AI

Needs, Challenges and Directions  
for Future Research





# Authors and acknowledgements

## Authors:

Jasminko Novak, European Institute for Participatory Media / University of Applied Sciences Stralsund

Kalina Drenska, European Institute for Participatory Media

Ksenia Koroleva, European Institute for Participatory Media

Lukas Pfahler, Technical University Dortmund

Lavinia Marin, Technical University Delft

Judith Möller, University of Amsterdam

Özlem Özgöbek, Norwegian University of Science and Technology

Martijn Willemsen, Eindhoven University of Technology

Dorothee Dersch, PD – Berater der öffentlichen Hand GmbH

Enny Das, Radboud University

Martha Larson, Radboud University

Katharina Morik, Technical University Dortmund

The work towards this report has been supported by the Volkswagen Foundation within its programme “Artificial Intelligence and the Society of the Future” as part of the project *Reflective AI: Understanding and Designing Environments for Reflective Information Practices in a Digital Society (2020 - 2021)* conducted by the European Institute for Participatory Media (Berlin), Radboud University (Nijmegen) and Technical University Dortmund.

The report would not have been possible without the interviews and workshop participation from experts and practitioners from diverse fields. We want to thank them for their invaluable contributions that helped us in shaping this report and developing further the notion of reflective AI.

We interviewed following experts for the purpose of this report: Maria Luciana Axente (PwC), Kirsti Elisabeth Berntsen (Norwegian University of Science and Technology), Xavier Brandao (#iamhere France), Randi Cecchine (University of Amsterdam), Dorothee Dersch (PD – Berater der öffentlichen Hand GmbH), Bart Goethals (University of Antwerp), Dietmar Jannach (Universität Klagenfurt), Lorenz Matzat (AlgorithmWatch), Sofia Papavlasopoulou (Norwegian University of Science and Technology), Nathalie Smuha (KU Leuven), Alex Urban (#iamhere Germany), Bob van de Velde (Nederlandse Publieke Omroep), Thabo van Woudenberg (Erasmus University Rotterdam) and Marcus Winter (University of Brighton).

Following researchers and practitioners took part in our interdisciplinary workshop “Reflective AI in a digital society” in May 2020: Mina Dennert (#iamhere International), Dorothee Dersch, (PD – Berater der öffentlichen Hand GmbH), Jana Diesner (University of Illinois at Urbana-Champaign), Virginia Dignum (University of Umeå), Jos Hornikx (Radboud University), Thomas Liebig (TU Dortmund), Lavinia Marin (TU Delft), Judith Möller (University of Amsterdam), Özlem Özgöbek (Norwegian University of Science and Technology), Niccolo Pescetelli (Max Planck Institute for Human Development), Till Plumbaum (Neofonie GmbH), Carina Prunkl (University of Oxford), Martijn Willemsen (Eindhoven University of Technology). Workshop organizers: Jasminko Novak (EIPCM, University of Applied Sciences Stralsund), Ksenia Koroleva (EIPCM), Kalina Drenska (EIPCM), Lukas Pfahler (TU Dortmund), Martha Larson (Radboud University), Enny Das (Radboud University) and Katharina Morik (TU Dortmund). We also thank Anne Janssen for her work in the project during her time at Radboud University.

Cite as: Novak, J. et al. (2021). *Towards Reflective AI: Needs, Challenges and Directions for Further Research*. European Institute for Participatory Media, Berlin, Germany, DOI: 10.5281/zenodo.5345643

Cover design: Ksenia Koroleva; Layout: Tjark Schlegel

This work is licensed under the terms of the Creative Commons Attribution License 4.0 which permits unrestricted use, provided the original author and source are credited. The license is available at: <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>



## Executive summary

**Ensuring a safe and responsible use of Artificial Intelligence (AI) cannot be solved alone through technological innovation and regulation, in spite of their importance.** The advantages of AI hide underlying problematic aspects, which can be harmful to users and need to be resolved to ensure a responsible and productive use of AI and its benefits. Research has already been addressing some of these problems, such as de-biasing AI to prevent discrimination, providing explanations of AI results, developing guidelines and certification mechanisms for trustworthy AI.

However, many of the problems connected to the use of AI technologies stem from **the lack of personal and societal experience with AI**. They mirror **not only the biases and inequalities reflected in the data and AI algorithms** but also **those from the organisational and societal contexts** in which AI is used and designed. To fully solve them, we need to understand **AI systems as socio-technical systems**: they are designed, built and used by people in different social contexts (e.g. individual, organisational, societal) that co-determine their interpretation and understanding, the nature of their use and the consequences thereof.


**How people conceive of AI, to what extent they understand its limitations, determines how they will perceive the results of AI systems and any possible consequences of their use.** In order to realize harm-free advantages of AI, it is necessary that we **cross the experience gap: both in private and professional use**. The experience gap is the difference between the experience that people have with AI on a day-to-day basis and the experience that they need in order to understand AI at the level necessary to harness its benefits and avoid its dangers.

In this report we **suggest a new framework for the development and use of AI technologies in a way that harnesses the benefits and prevents the harmful effects of AI**. We name it **Reflective AI**. The notion of Reflective AI that we propose calls for adopting a holistic approach in the research and development of AI to **investigate both what people need to learn about AI systems to develop better mental models** i.e. an **experiential knowledge of AI**, to be able to **use it safely and responsibly**, as well as **how this can be done** and supported.

The Reflective AI framework describes three main levels where interventions are needed: **end-users, AI developers and designers, AI regulators**. For **end-users**, a better understanding of key properties of AI is in the centre of the framework. To achieve this, solutions that allow and support **experiential learning about key properties of AI** that are normally hidden from users need to be developed. Regarding **AI developers and designers**, the framework is concerned with what they need to understand about user needs and what changes in their work practices are required to be able to support the end-users better in achieving reflective AI use. At the level of **AI regulators** the framework highlights the challenge of how public policies could support the development of a better understanding of AI among end-users.

Implementing a transdisciplinary and participatory approach that involved researchers and societal actors from different areas, the following main observations for further research and practice towards the vision of Reflective AI were identified:

**1) Enabling people to understand AI and the consequences of its use and design is more challenging than previously thought.** The risks of AI stem not only from problematic technological designs, but also from the lack of awareness of end-users and societal stakeholders about consequences of an uncritical application of AI and unquestioned reliance on its results.



**2) AI needs to be demystified in order to overcome the experience gap and reach AI literacy to ensure productive and responsible use.** Future research needs to better understand misconceptions of AI and the AI experience gap and find solutions to overcome them.

**3) AI models need to be interpretable by design. Interpretability of AI is a prerequisite for reliable explanations and reflective use of AI by end-users, developers and designers alike.** Research on interpretable machine learning combined with human-AI interaction and AI ethics is crucial for the development of trustworthy AI systems that are verifiable by experts and whose workings and consequences can be appropriately explained to lay end-users and stakeholders.

**4) Designing for Reflective AI experiences requires changes in work practices of AI developers and designers.** Future AI development should be more interdisciplinary by definition. User experience design should make inherent properties and risks of AI models visible (e.g. sensitivity, diversity, privacy), without overburdening the users. Educating user experience designers is crucial, as their work shapes the perceptions and use of AI systems.

**5) Reflective adoption of AI innovations in organisations requires changes in organisational values, value chains and processes to align with the needs of different actors.** Apparent trade-offs between commercial goals, the values of the users and the principles of transparency, fairness and explainability need to be consciously resolved by reconsidering company values and commercialization models. This requires participative processes that address the interdependencies and enable dialogue between different actors (e.g. employees and managers, AI developers and AI users). Establishing organisational laboratories for Reflective AI experiences can facilitate organisational learning about AI and its potentials for the organisation.



# Table of Contents

|  |           |
|--|-----------|
| <b>Authors and acknowledgements</b>  | <b>1</b>  |
| <b>Executive summary</b>   | <b>2</b>  |
| <b>1. Introduction</b>   | <b>6</b>  |
| 1.1 Purpose and goals of the report  | 7         |
| <b>2. What is Reflective AI and why is it needed?</b>                                  | <b>9</b>  |
| 2.1 The risks and harms of unreflected use of AI                                       | 9         |
| 2.2 Main research perspectives on ensuring a safe and responsible use of AI            | 12        |
| 2.3 The need for a Reflective AI   | 14        |
| <b>3. What do people need to understand about AI to use and govern it responsibly?</b> | <b>17</b> |
| 3.1 End-users & broader public   | 17        |
| 3.1.1 Demystifying AI  | 17        |
| 3.1.2 Operational principles and hidden properties of AI                               | 19        |
| Sensitivity  | 19        |
| Temporal effects   | 19        |
| Non-linearity  | 20        |
| Birds-eye view   | 20        |
| Privacy preservation   | 20        |
| 3.2 AI developers and designers  | 21        |
| 3.3 AI regulators  | 22        |
| <b>4. How can we design systems and solutions that support a reflective use of AI?</b> | <b>24</b> |
| 4.1. Transparency of AI presence (“AI inside”)   | 25        |
| 4.2 Understandability of operational principles, properties and risks of AI            | 27        |
| 4.2.1 Explaining operational principles  | 27        |
| 4.2.2 Enabling users to learn about key properties of AI                               | 29        |
| Example approach: Experiential learning environments for Reflective AI                 | 31        |
| Example approach: Design issues for Reflective AI in recommender systems               | 33        |
| 4.3 Control over the use of personal data in AI (“privacy preserving AI”)              | 34        |
| <b>5. Work practices in AI design, organisational and structural changes</b>           | <b>37</b> |
| 5.1 (New) work practices of AI designers and developers                                | 37        |
| 5.1.1 Supporting user experience designers in learning about AI                        | 38        |
| 5.1.2 Integration of ethical awareness into AI development and teaching                | 38        |
| 5.1.3 Integrating interdisciplinary approaches to consider context of use in AI design | 39        |



|   |           |
|---|-----------|
| Case study 1: Addressing the problem of misinformation by considering the context in which communication occurs on social network sites | 40        |
| Case study 2: Accounting for user-specific factors when providing behavioral change recommendations                                     | 41        |
| 5.2 Organisational practices for Reflective AI  | 42        |
| 5.2.1 Integrating reflective AI in organisational innovation adoption   | 42        |
| 5.2.2 Value changes of commercial organisations   | 43        |
| 5.3. Structural changes for Reflective AI   | 44        |
| 5.3.1 Auditing and control of algorithm development and deployment  | 45        |
| 5.3.2 AI literacy and public education about AI   | 45        |
| <b>6. Directions for further research</b>   | <b>47</b> |
| 6.1 Demystifying AI: Transparency, Understandability, Diversity, Control  | 49        |
| Transparency of AI presence (“AI inside”)   | 49        |
| Understandability of operational principles, properties and risks of AI   | 50        |
| Diversity and “birds-eye view”  | 52        |
| Control over the use of personal data in AI (“privacy preserving AI”)   | 53        |
| 6.2 Designing for experiential learning and reflective AI experiences   | 54        |
| 6.3 Work practices in AI design & development   | 56        |
| 6.4 Organisational adoption of AI   | 57        |
| <b>7. Summary</b>   | <b>58</b> |
| <b>References</b>   | <b>59</b> |



# 1. Introduction


AI is increasingly used by online platforms and systems that are part of our daily lives. It plays a growing role in determining how we access and consume information, how we make judgements based on it and how we interact and perceive each other. AI promises great benefits for dealing with complex situations and for enhancing human cognition. **A productive and responsible use of AI promises many benefits, from better medical therapies and decision-making in complex situations, to safer traffic, fighting climate change and supporting sustainability, to fostering creativity and learning,** to name but a few. However, there has been an increasing awareness that the advantages of AI also hide underlying problematic aspects, which can be harmful to users and that need to be resolved to ensure a responsible and productive use of AI.

**AI systems and technologies have important limitations and these require careful consideration in the design and use of AI.** AI is data-driven but designed, built and used by people: as individuals, as organisations and as society as a whole. All of these are sources of “imperfections”. Data can be incomplete, unrepresentative and biased. People, organisations and societies can be biased, unfair and discriminating in their behaviour, decisions and beliefs.

It is no news anymore that these problematic aspects have found their ways into AI systems we build and use. They are **sources of problems that can cause societal harm and prevent a productive and beneficial use of AI.** AI systems have been found to mirror existing historical, cultural, gender, economic and political inequities (e.g. Bolukbasi et al., 2016; Lambrecht & Tucker, 2019), unless explicitly designed not to do so. Deep learning has been criticized for inducing a false sense of certainty in the accuracy of its results (Guo, 2017; Buschjäger et al., 2020). The use of AI can intensify discriminatory practices (Dastin, 2018; Raghavan et al., 2020; Hill, 2020) or reinforce existing human biases such as confirmation bias (Nickerson, 1998) and social phenomena such as herding (Michael & Otterbacher, 2014; Raafat et al., 2009) and echo-chambers (Garrett, 2011; Quattrociocchi et al., 2016). This can intensify polarization of the public discourse (Adamic & Glance, 2005; Del Vicario et al., 2016; Del Vicario et al., 2017) and contribute to the spread of online manipulation and misinformation (Del Vicario et al., 2016; Vehof et al., 2019). Such potential harms of AI pose a fundamental challenge to democratic societies because they can decrease trust in fair treatment and in the transparency of democratic processes.

Research has already been addressing some of these problems in different ways: de-biasing AI to prevent discrimination (Raghavan et al., 2020), providing explanations of AI results (Abdul et al., 2018; Biran & Cotton, 2017), creating guidelines and certification mechanisms for trustworthy AI (AI HLEG, 2019; Brundage et al., 2020). But **many of these problems cannot be solved purely technologically, as they also stem from the lack of personal and societal experience with AI and from the biases of social contexts in which AI is designed and used.** To fully address them, we need to understand AI systems as socio-technical systems. Systems that are **designed, built and used by people in different social contexts** (e.g. individual, organisational, societal) that co-determine their interpretation and understanding, the nature of their use and the consequences thereof.

**How people conceive of AI, to what extent they understand its limitations, strongly determines how they will perceive the results of AI systems and any possible consequences of their use.** It is not only the general public that often relates AI to a “mystical” intelligence from SciFi movies, unaware that AI is present in many daily activities they perform, such as browsing on the Internet or in the feeds of their social networks. **Misconceptions about the nature and the behaviour of AI systems are also held by decision-makers or policy-makers when they make decisions that affect individuals and society alike .**



This is largely due to the complex and hidden properties of AI behaviour that are neither readily observable nor easily understandable for people, while influencing the effects of AI on individuals and society (e.g., radicalization on YouTube (Kaiser & Rauchfleisch, 2018; Ribeiro et al., 2020), the rabbit hole effect (O’Callaghan et al., 2015), privacy risks (Larson et al., 2017), health and public safety (Whittaker et al., 2018)).

What the data-driven and probabilistic nature of AI technologies imply for their results and the unintended effects of their use is hard to intuitively understand. The misconceptions of AI and the lack of an underlying understanding of the behaviour of AI systems lead to wrong expectations and unreflected use. This threatens the productive use of AI to the benefit of individuals, organisations and the society as a whole.

The notion of *Reflective AI* therefore calls for the investigation and development of new approaches that can enable a more **reflective use and design of AI that empower people and the society at large to harness the benefits and avoid the potentially harmful effects of AI.**

Addressing this challenge requires novel approaches that acknowledge but go beyond the existing technological solutions (e.g. explainability, de-biasing, fairness, trustworthy AI) by understanding **AI systems as socio-technical systems** and by **increasing the capabilities of people and societies to productively reflect on the nature and consequences of their use of AI.**

We thereby understand the term of *Reflective AI* as a broad umbrella connecting different challenges and research directions that are required to reach its goals. Some of the guiding questions that have informed our initial conception of the problem and solution space of *Reflective AI* include (but are not limited to):


1. How can we enable people to develop an appropriate *experiential* understanding of AI that enables them to reflect on their use of AI and its personal and societal impact?
2. How can we design environments that encourage critical reflection on the behaviour of AI systems, their results and the information they mediate?
3. What else is needed so that *Reflective AI* effectively leads to more responsible use of AI allowing people and societies to harness its benefits and prevent harm?
4. What normative understandings and problems from the social, ethical and democratic perspectives should be considered when defining the notion of reflective information processing and enabling *Reflective AI* solutions?

## 1.1 Purpose and goals of the report

This report seeks to map out a variety of perspectives from different scientific disciplines, research areas and societal actors, as to what constitutes the main problems and challenges, possible solution approaches and promising research directions for the idea of *Reflective AI*.

The wide scope of our notion of *Reflective AI* is deliberate. It seeks to provide a broad frame of orientation that can help relate and connect the many different disciplines and research areas whose contributions will be required to address this challenge that is transdisciplinary by its very nature. Instead of defining the problem in terms of the perspective and knowledge of a specific discipline, we ask: **what perspectives and knowledge need to be brought together to understand and successfully address the challenges that are highlighted by the notion of *Reflective AI*?**





Against this background, this report presents the insights and findings of the planning grant project *Reflective AI* funded by the Volkswagen Foundation and of its outreach to a broader community of researchers, practitioners and societal stakeholders.

The original project grant involved three partners: the European Institute for Participatory Media, Radboud University and the Technical University Dortmund. However, in order to expand the range of perspectives the project has reached out to a broader research community and societal stakeholders.

In an online workshop “*Reflective AI in a digital society*” in May 2020 we brought together researchers and practitioners from academia and industry from a wide range of fields: from Artificial Intelligence and Machine Learning, HCI and Interactive Systems to Computational Social Science, Communication Science, Education and Philosophy. This was accompanied by a series of expert interviews to elicit views and insights from even a broader range of practitioners and stakeholders from public organizations and companies, online media platforms and journalists, schools and universities, and from specific fields of research (e.g. AI literacy, human-centered AI).

Workshop participants have been invited to contribute to parts of this report and those who have provided such contributions have been included as co-authors. Participants who didn't provide contributions to the report directly, but participated in the workshop have been acknowledged as workshop participants. All experts and stakeholders who took part in the interviews and reviewed the report have also been acknowledged in the list of consulted experts.

This report thus synthesizes the main findings from this explorative and collaborative, transdisciplinary process to map out the theme and research directions of what we see as an emerging field of *Reflective AI*. We hope that this can provide an impulse for new approaches in research and practice on achieving the vision of empowering a responsible use and design of AI that harnesses its benefits and avoids potential harm.



## 2. What is Reflective AI and why is it needed?

*This chapter describes and motivates the notion and vision of Reflective AI in more detail and from different perspectives. What are the main problems and challenges it addresses and why is it needed?*

The attention to the challenge of ensuring that AI technologies are used in a safe and responsible way that prevents harmful individual and societal effects is not new. Already in early AI research, societal and ethical issues have been pointed to: e.g. from the expectations and premises associated with different visions of artificial intelligence (Weizenbaum, 1976; McCarthy, 1979; Versenyi, 1974; Pană, 1973), to explainability of expert systems (Clancey, 1983), to social implications and ethical challenges in specific domains (e.g. Boden, 1978; Szolovits & Pauker, 1979; Lusted, 1978; Croy, 1989).

More recently, a number of research perspectives have been formulated that emphasize different challenges and solution approaches to ensuring a safe and beneficial use of AI in society. This research has been referred to under many different themes and approaches, from **Responsible AI** (Dignum, 2017; Fjeld et al., 2020) to **Explainable AI** (see reviews in e.g. Arrieta et al., 2020; Biran & Cotton, 2017; Abdul et al., 2018; Langer et al., 2021) and **Trustworthy AI** (AI HLEG, 2018; Chatila et al., 2021; Brundage et al., 2020), to most recently **AI Literacy** (Long & Magerko, 2020).

Our notion of **Reflective AI** shares the underlying concerns and some premises of these perspectives but it also differs in a specific focus that we see as underrepresented. In the next sections we first review common risks and harms of an unreflected use of AI and the approaches of the above perspectives on ensuring a safe and responsible design and use of AI. In doing so we highlight the relation to and differences to our notion of Reflective AI.

### 2.1 The risks and harms of unreflected use of AI

In the last decade there has been a rising awareness about the advantages of AI hiding underlying problematic aspects, which can be harmful to users as individuals and the broader society alike.


This starts already with what one could consider mundane daily activities which people perform without a second thought. For instance, many of our everyday actions are supported by recommender algorithms predicting what music we like, which shows to watch, what news feeds to read and what items to shop next (Konstan & Riedl, 2012a,b). Such recommender systems are effective AI tools that help users to overcome information overload, though some worries have been voiced that they might lead to filter bubbles (Pariser, 2012) by intransparently limiting the content and information to which users are exposed.

Moreover, as business models of online companies are often based on captivating users to spend as much time as possible with their content, the design of such algorithms can be biased towards artificially keeping users attention, not aligned with the actual value for the user (e.g. so-called clickbaiting (Potthast et al., 2016)). This might also occur inadvertently, for example, as Neil Hunt argued in his keynote at REcSys 2014 the Netflix's otherwise effective recommendation algorithm might in some cases actually be reinforcing binge watching rather than adding value for the user<sup>1</sup>.

Perhaps even more pressingly from the perspective of societal consequences, AI systems can reinforce existing human biases such as confirmation bias (Nickerson, 1998) and social phenomena such as herding (Michael & Otterbacher, 2014; Raafat et al., 2009) and echo-chambers (Garrett, 2009; Quattrociocchi et al., 2016). In this context, echo-chambers are defined as ideologically homogeneous online spaces of like-minded individuals where people reinforce each other's beliefs which results in attitude polarization (Adamic & Glance, 2005; Del

---

<sup>1</sup> <https://youtu.be/IYcDR8z-rRY> (from 56:00 on)



Vicario et al. 2015; Del Vicario et al., 2017). The idea of echo chambers is based on two main components: 1) algorithmic curation through which people only get recommendations for types of information they have previously engaged with and/or liked and 2) selective exposure - a behavioral aspect that points towards the tendency among people to group together with like-minded others (Cardenal et al., 2019; Wollebaek et al., 2019). Some scholars have pointed out that echo chambers threaten a healthy public life by increasing group polarization (as echo chambers are devoid of attitude-challenging content, Bakshy et al., 2015), audience fragmentation and the circulation of fake news (Cardenal et al., 2019).

YouTube is a prominent example of a social network where AI recommendations can push users further down the “rabbit hole” of right wing radicalization (O’Callaghan et al., 2014). Ribeiro, Ottoni, West, Almeida and Meira (2020) investigated the so-called radicalization pipeline on YouTube by analysing over 300,000 videos from channels of the Intellectual Dark Web, Alt-Lite and Alt-Right. They found that these three groups increasingly share the same user base, that users migrate from milder to more extreme content (users that initially comment only on IDW or Alt-Lite content later comment on Alt-Right content), and that alt-lite content is easily reachable from IDW channels and alt-right through both IDW and alt-lite channels through recommendations. Through examples like this we see how behavioral patterns and cognitive biases could be reinforced through the use of AI technologies, which - especially when aggregated on a massive scale - can contribute to the development of extremist beliefs that are harmful for democratic societies and public discourses.

Furthermore, recommender systems have been shown to mirror existing historical, cultural, gender, economic and political inequities (e.g. Bolukbasi et al., 2016; Lambrecht & Tucker, 2019), while deep neural networks have been criticized for inducing a false sense of certainty in the accuracy of their results (Guo et al., 2017; Buschjäger et al., 2020). The combination of these two characteristics of AI technologies has been shown to have severe individual and societal consequences, such as the intensification of discriminatory practices in recruitment processes. In such scenarios AI algorithms might not necessarily recommend the most skilled candidates, but rather candidates that fit the profile of people who have historically been more often employed at a given company or position (e.g. men rather than women in the IT sphere) (Dastin, 2018; Raghavan et al., 2020).


Racial and class inequalities rooted in historical data used for training recommendation algorithms have already affected the access of people to medical health care (Strickland, 2019) even when algorithms were specifically created to not take race into considerations in order to avoid precisely such biases. Recommender algorithms could furthermore be biased when assessing the defendant’s future risk for misconduct in the criminal justice system (Chohlas-Wood, 2020), while incorrect results of facial recognition software have already led to charging innocent people with crimes they didn’t commit (Hill, 2020).

A particularly problematic aspect arises when facial recognition AI technologies are based on the pseudoscientific and very questionable theory of physiognomy – the notion that based on the physical appearance of a given individual, conclusions could be drawn about their personality, inner characteristics, sexual and political orientation etc. (for an overview see e.g. Bendel, 2018; Fernández-Martínez & Fernández, 2020). Such “predictions” about an individual based on their looks are also proven to be deeply racist in their origins (e.g. Belting, 2013; Campe & Schneider, 1996), nonetheless both commercial<sup>2</sup> and research projects<sup>3</sup> claim to have developed algorithms that can tell whether someone is aggressive or a criminal solely by analysing their facial

---

<sup>2</sup> <https://www.facepotion.com/>

<sup>3</sup> See the controversy around the research paper “Automated Inference on Criminality Using Face Images” (2016) by Xiaolin Wu and Xi Zhang of the Jiao Tong University in Shanghai.



appearance. Some companies<sup>4</sup> are using facial recognition technologies and insist on being able to assess personality characteristics of job applicants such as their openness, conscientiousness, extraversion, agreeableness, and neuroticism based on their appearance in video materials created for the recruitment process. As experiments have shown, the results of such algorithmic assessments of human behavior can be influenced by factors such as whether or not the applicant wears glasses or a headscarf, the brightness of their video, or even objects in their background<sup>5</sup>.

These developments show that sensibilizing experts from different domains about the risks of relying on AI recommendations without an understanding of and a critical reflection on how such recommendations are produced and what ethical considerations should be taken into account when designing (or deciding not to design) AI applications is a critical step in ensuring that AI is used and developed responsibly.

Moreover, governments worldwide are increasingly relying on automated decision making systems in domains such as immigration (Akhmetova, 2020) and allocation of resources such as social, welfare and child care benefits (e.g. Henley, 2021). However, such systems are often developed by private companies and not undergoing sufficient testing and controlling processes before being implemented (Richardson et al., 2019), thus often resulting in discrimination against already marginalized societal groups when it comes to access to public resources (e.g. Geiger, 2021; Lecher, 2018).

Finally, the advances in the development of AI technologies put a strong focus on concerns surrounding the breach of individual user privacy, the surveillance capacities of such technologies and the possible implications for civil liberties (e.g. Whittaker et al., 2018). Techniques that “analyze video, audio, images, and social media content across entire populations and identify and target individuals and groups” (Whittaker et al., 2018: 12) are used by private actors and governments alike for large-scale data collection, while users are rarely aware of the fact that such data is being collected.

As such, AI could pose a fundamental challenge to democratic societies by decreasing trust in fair treatment and in the transparency of democratic processes. The question of auditing and controlling the development and implementation of AI technologies, as well as the question of training public servants to understand better, not overtrust and be able to audit AI-based decision making systems is thus ever more pressing.

With the growing awareness of such problems in the AI research community many shortcomings of current AI designs are being addressed in research (e.g. de-biasing datasets and algorithms (Raghavan et al., 2020), developing fairness models for AI (Zhang et al., 2020), providing explanations of AI results (Sokol & Flach, 2018), certification mechanisms for AI algorithms (Kulesza et al., 2013; Normann, 1983).

However, rather than being solvable through technology alone, both harnessing *benefits* and *preventing potential harms* of AI depends on a complex interplay between technology, individual behaviour, organizational and societal dynamics and governance. As the above examples illustrate, **the risks and harms of AI can stem both from problematic technological designs, as well as from the lack of awareness of end-users and societal stakeholders about potential consequences of an uncritical application of AI and unquestioned reliance on its results.**

---

<sup>4</sup> <https://www.retorio.com/>

<sup>5</sup> For more information see the investigative project of BR24: <https://web.br.de/interaktiv/ki-bewerbung/en/>



## 2.2 Main research perspectives on ensuring a safe and responsible use of AI

Against this background, various perspectives have been formulated that emphasize different challenges and solution approaches to ensuring a safe and beneficial use of AI in society. The notion of **Responsible AI** has developed into an umbrella term for describing guiding principles that should be adhered to in order ensure a “safe, beneficial and fair use of AI technologies to consider the implications of morally relevant decision making by machines, and the ethical and legal consequences and status of AI” (Dignum, 2017: 4698).

While different authors and societal actors (e.g. research and academia, companies, NGOs, governments) have proposed somewhat different governance frameworks for ensuring a safe and responsible use of AI they all tend to share the emphasis on ensuring that the design, implementation and use of AI considers ethical aspects in accountable and transparent ways and that it is aligned with moral, societal and legal values (e.g. Dignum, 2017; Telefónica, 2018; Rao et al., 2019; Eitel-Porter et al., 2021).

The findings of a recent study (Fjeld et al., 2020) of 36 different published frameworks suggest that meanwhile a consensus has emerged around a shared set of guiding principles for Responsible AI that include: **privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human-control of technology, professional responsibility, promotion of human values.**


The work on ensuring transparency and explainability of AI systems under the umbrella of **explainable AI** (Arrieta et al., 2020; Biran & Cotton, 2017; Abdul et al., 2018; Langer et al., 2021) directly relates to supporting a responsible design and use of AI by investigating how AI systems and their results can be made more explainable or interpretable for different types of users (see e.g. Wang et al., 2019 for an overview).

Thereby, a number of research contributions have focused on the technical aspects of explaining the reasons behind the results of complex AI algorithms that are difficult to understand for non-experts. More recently, explainability research has been more specifically motivating the desired types of explainability with the requirements related to the principles of responsible AI (e.g. Rudin, 2019; Arieta et al., 2020; Langer et al., 2021).

Introducing explainable AI in organizations currently tends to be motivated by legal accountability (e.g. Bhatt et al., 2020) and can help implement safeguards for non-discrimination and fairness, e.g. by making it easier to interpret and assess system behaviour, which can in turn facilitate more conscious design and implementation practices (ibid.). The underlying assumption of explainable AI is that by making results and (sometimes) the functioning of AI algorithms explainable and interpretable to users, this can make the use of AI safer. Explanations are expected to increase the capacity of the users to correctly interpret the meaning of AI results, assess their reliability and take decisions that are aligned with ethical, organizational and legal requirements.

**Trustworthy AI** aims at ensuring a safe and responsible use of AI by making it verifiable that AI systems actually adhere to their stated goals, values and overall principles of responsible AI. This can occur through methods and mechanisms that developers themselves can apply to describe and verify “claims about AI development, with a focus on providing evidence about the safety, security, fairness, and privacy protection of AI systems” (Brundage et al., 2020: 1).

Moreover certification approaches are being pursued that describe which properties of AI systems should be certifiable (e.g. fairness, transparency, reliability, safety, privacy), how this could be achieved and communicated (e.g. through certification labels) to ensure trustworthy AI implementations (Chatila et al., 2021; Cremers et al., 2019).



Most recently, attention has been developing towards **another part of the equation that has received little attention: what would users need to know in order to use AI effectively, safely and with a critical mind? And how can we support end-users learning what they need to know about AI to achieve that (Long & Magerko, 2020)?** These questions are at the core of our notion of Reflective AI.

Existing work addressing these questions has so far been relatively rare and scattered. It has mostly focused on different forms of education approaches that aim at teaching the basics of AI to non-technical audiences<sup>6</sup>, underrepresented audiences<sup>7</sup> or school children (e.g. Zimmer, 2018; Druga et al., 2019; Khan & Winters, 2017). In order to inform the development of suitable approaches, some HCI research has been increasingly looking into how people conceive of and make sense of AI from the perspective of explainability (Abdul et al., 2018).

The most comprehensive approach up to date is a recently proposed conceptualization of **AI literacy** “as a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace” (Long & Magerko, 2020: 2). It proposes an initial set of competencies that people should acquire to become AI literate, derived from an extensive literature review. It also provides a set of recommendations for AI developers on how to incorporate these considerations into the design of AI systems. This highlights one area that has so far received little attention in the existing approaches under the umbrella of responsible AI, explainability and trustworthy AI.

Our notion of Reflective AI could thus be considered as a specific perspective on the broader concept of AI literacy. **The guiding questions and goals of AI literacy are also at the core of the concept of Reflective AI. However, we see them as a “missing link” between the guiding principles and regulatory guidelines of responsible AI, the efforts at making AI more explainable and the certification mechanisms of trustworthy AI.**

In addition to the closely related goals and questions, the perspectives of AI literacy and Reflective AI share some of the envisioned competencies (e.g. “Recognizing AI”, “Understanding AI strengths and weaknesses” (Long & Magerko, 2020)). However, the Reflective AI approach differs in two main ways. First, we focus more specifically on **what exactly the users should be able to critically assess about AI**: e.g. understand potential individual and societal harms and what they result from. Second, it differs in defining **what it is that people would need to understand about AI (e.g. hidden properties of AI) in order to be able to productively reflect on its use and effects**.

Perhaps the biggest difference is that the proposed set of 16 competences for AI literacy seems geared toward the notion of competences as commonly found in formal academic education: e.g. “Competency 7 (Representations) - Understand what a knowledge representation is and describe some examples of knowledge representations” or “Competency 9 (ML Steps) - Understand the steps involved in machine learning and the practices and challenges that each step entails” (Long & Magerko, 2020: 6).

In contrast, the notion of *Reflective AI* emphasises the need to develop an *experiential understanding of what constitutes the special nature and properties of AI*, what kind of individual and societal implications (e.g. harms) they can carry and what that implies for ensuring a safe and responsible use of AI both for individuals and the society as a whole.

---

<sup>6</sup> A prominent example would be the international course *Elements of AI*: <https://www.elementsofai.com/>

<sup>7</sup> Such as the initiatives *AI4All* ( <https://ai-4-all.org/> ) or *Ready AI* (<https://www.readyai.org/>)





## 2.3 The need for a Reflective AI

Our notion of Reflective AI calls for the investigation and development of new approaches that enable a more reflective use and design of AI that empower people and the society at large to harness the benefits and avoid the harmful effects of AI.

We propose that in order to achieve that, in addition to the concerns and principles of the existing approaches to responsible use and development of AI, it is necessary that we cross the *experience gap*. The experience gap is **the difference between the experience that people have with AI on a day-to-day basis and the experience that they need in order to understand AI at the level necessary to enjoy its benefits and avoid its dangers.**

Why does this experience gap (still) exist? The reasons are manifold. To start with, in spite of a widespread presence of AI in professional and everyday life it is still difficult for people to both recognize the use of AI in the different systems, and to understand the implications thereof (Eslami et al., 2019; Eslami et al., 2015). **Systems using AI often don't present themselves as such and the consequences of that for what they do.**


Historically, the underlying principles, properties and behaviour of AI are much different from digital systems people have become accustomed to. **The probabilistic nature of AI mechanisms and the consequences of that compared to more deterministic systems are hard to fathom.** The much discussed **intransparency of many AI systems and algorithms ("black boxes")** causes further difficulties for users to understand the nature of systems they are dealing with.

**As a result, people form misconceptions of both AI as such, as well as of systems in which AI is used in ways not directly discernable for them or that are too complex to be understood without technical knowledge** (Eslami et al., 2019; Burrell, 2016).

For example, many ubiquitous online platforms are often perceived as platforms for information access, content sharing or social interaction (e.g. Google, YouTube, Facebook) without an awareness of the underlying AI algorithms and their implications (Eslami et al., 2015). **This makes it difficult for people to correctly "categorize" their experiences with such systems and leads both to the lack of prompts for the necessity to reflect on their use and to the lack of support to do so.**

Although awareness is growing about the need to alert the users about the presence of AI (see e.g. Fjeld et al., 2020), the implementation of this requirement in the design and provisioning of AI systems in practice is still far behind. This is further aggravated by the widespread tradition of "seamless design" of interactive systems that hides the complexity and underlying system mechanics from users as a premise of a frictionless and enjoyable experience (Hamilton et al., 2014; Weiser, 1994). Although the appropriateness of this paradigm and its potentially harmful consequences have been questioned in HCI research itself (Inman & Ribes, 2019; Hamilton et al., 2014), the seamless design tradition remains largely unchallenged in business practice.

More importantly, while a large body of research on explainable AI has investigated possibilities for explaining the reasoning of AI systems and the results they produced to users, existing approaches largely assume that this can be achieved without understanding the underlying fundamental principles and properties of AI itself. Another view is that while the reasons for specific AI results might be explainable or even directly interpretable (Rudin, 2019), the underlying workings of the employed AI models cannot be explained because they are too complex for non-experts to understand.



A key challenge that we see is that there are **fundamental principles and properties of AI that need to be understood by users of AI systems in order to form an appropriate image (a *mental model*)** of the system they are using and thus appropriately understand the nature of its outputs. The crucial problem is that these fundamental properties of AI are **commonly hidden from users and cannot be directly experienced via casual interaction**.

For example, many AI methods are based on complex statistical models and **probabilistic reasoning and involve non-linearity and uncertainty, phenomena that are difficult to grasp and understand intuitively for non-experts**. Many AI methods are sensitive to minor variations of input that can lead to big changes in the results. This can lead to **misplaced trust in the reliability of AI results** - that is difficult to fix with individual explanations without an underlying awareness of the extent of their importance. The effects of AI also accrue over time and at large scale, often through gradual changes that are not directly perceptible for users (e.g. changes in attitudes due to exposure to recommendations of specific content).

**Since most people only experience a small fraction of the behaviour of an AI system that tends to be highly dependent on users preference profiles and patterns of interaction, it can be difficult to perceive or understand potential harms caused by their indiscriminate use** (e.g. how recommender systems can lead to radicalization or exacerbate polarization) .

The ability of AI to protect users, for example, in their **privacy**, is also not directly observable. This leads to wrong assumptions e.g. about the inevitability of surrendering large amounts of personal data as a condition for system use. This directly constrains the possible realizations of the principle of autonomy for the users of such systems.


Moreover, there is an inherent trade-off between conscious **effort needed by users to actively analyse and reflect on the behaviour of a system in use, compared to efficiently achieving their purpose** (e.g. finding information, taking a decision, being entertained). Existing approaches to explainability largely focus on *static* explanations that aim to explain how *a given* system has produced a *specific* result (Adabi & Berrada, 2018). But this cannot adequately support the understanding of essential properties of AI systems, the lack of which aggravates many of the observed negative personal and societal effects of indiscriminate use of AI and hampers its responsible uptake and beneficial use (see Section 2.1).

**Due to the *lack of possibilities and occasions to experience and reflect on the main properties of the behaviour of AI systems and the consequences thereof, few people have thus developed appropriate mental models of AI systems.***

What mental models people have of AI and how these are constructed is still not well researched, although the work on these issues is picking up (e.g. Hernandez-Bocanegra & Ziegler, 2021; Alizadeh et al., 2021). However, little work has yet been done on how the development of more suitable mental models could be supported - including the possible consequences of the existing misconceptions.

Since most people lack suitable mental models of AI systems, an overall idea of how AI systems work and of their possible personal and societal impacts (a kind of *experiential knowledge* of AI), they are unable to critically assess their results and reflect on the effects of their indiscriminate use. This makes it not only difficult to develop a more conscious, reflective practice in their use of AI, but also decreases their ability to act as responsible citizens e.g. by weighing online information, making informed judgments and counteracting the polarization of online communication.





Making AI systems understandable for laypersons is particularly difficult due to the nature and complexity of underlying algorithms that are often difficult to interpret and understand even for AI experts. However, we argue that people do not need to achieve expert-level understanding of AI, but an experiential understanding of its essential principles and properties. Such an understanding of AI would allow people to decide for themselves which role they allow AI to play in their personal lives. Informed citizens are necessary in order to participate in the required civic discourse about governmental regulations of AI.

**The notion of *Reflective AI* that we propose asks us to adopt a holistic approach regarding both *what* people need to learn about AI systems to develop better mental models i.e. an experiential knowledge of AI and to be able to use AI safely and responsibly, as well as *how* this can be done and supported.**

It emphasises that while important, it is not enough to provide people with notifications about the presence of an AI system, the explanations of its results and information about purely functional affordances of AI technologies. Rather, we propose that there is a great need for enabling people to develop an understanding of key principles and properties of the ways in which AI systems operate and to be empowered to reflect on potential personal and societal implications of the use of AI in different contexts.

However, at the same time, as researchers and designers of AI systems we need to better understand what makes it difficult for people to develop this kind of understanding and capacity for reflective use. **We need to better understand what should constitute this kind of understanding: what should people know and understand about AI in order to be able to enjoy its benefits and avoid harms?** And we need to find out how we can design AI systems, learning environments or interventions that provide opportunities for people to develop such kinds of understanding.

In line with the overall approach of Responsible AI, such a notion of Reflective AI recognizes that ensuring this cannot be achieved by focusing alone on the end-users and researchers. Rather it requires the awareness, action and collaboration of different actors at different levels of society, beyond education and research. Companies that apply, develop, implement, and provide AI also carry the responsibility for addressing these needs and challenges in the design and provision of their products and services.

From the perspective of Reflective AI this also calls for regulatory frameworks to make sure that people using AI can have the *occasions and means to experience and reflect* on the properties and effects of the behaviour of AI systems (e.g. obligatory training courses for specific areas of AI application) in order to support a reflective use that can prevent personal and societal harms.



## 3. What do people need to understand about AI to use and govern it responsibly?

*This chapter discusses the public perception of AI technologies, (mis)conceptions and concerns about AI that can hinder its reflective and responsible use. It focuses on the main needs that should be addressed in order for people and communities to be able to harness the benefits and avoid the negative effects of AI technologies.*

We propose that **increasing a reflective use of AI can only be successful as a joint effort, a shared responsibility, between the designers and developers of AI algorithms and systems that use them, the companies and organisations that employ or provide such systems, the end-users and (inter-)governmental actors providing the required regulatory frameworks.**

Accordingly, we distinguish between three different levels of analysis and types of actors throughout the chapter: end-users (individuals or the general public broadly), AI developers and designers (in companies, organisations and research) and those responsible for the regulation of AI technologies (states, public institutions, supranational structures).

Each of these groups of actors has a different level of responsibility when it comes to the outcomes of AI technologies and needs to overcome different problems when dealing with AI. The chapter summarizes insights from existing literature and research on the topic, as well as the results from expert and stakeholder interviews conducted in the course of the *Reflective AI* project.


### 3.1 End-users & broader public

Understanding public and end-user perception about AI technologies is important for two main reasons. On one hand, public concerns about AI can translate into regulatory activity with potentially serious implications (AI100, 2016). But also (mis)conceptions about what (existing) AI technologies are capable of could lead to user neglect of already existing risks of using AI technologies such as overtrusting the AI decision-making processes (Howard, 2020), data security breaches, creation of echo-chambers, filter bubbles and similar. Even if the topic is of high relevance, there are surprisingly few empirical studies or research on the public perception of AI technologies and most of the available empirical data comes from polls that measure recent attitudes towards AI technologies (BSA, 2016; 60 Minutes/ Vanity Fair Poll, 2016).

Despite the sharp increase in discussions on AI in popular media outlets since 2009 and the overall more optimistic public perception about such technologies (Fast & Horvitz, 2017), there is an ongoing trend outlining specific concerns that people have such as the fear of loss of control of AI (ibid.), ethical consideration about the lack of ability of AI technologies to integrate moral judgements in the decision-making processes (ibid.) and the fear of job losses to AI in the near future (BSA 2015).

#### 3.1.1 Demystifying AI

These and similar empirical findings were echoed in the expert and stakeholder interviews conducted within the *Reflective AI* project. The majority of the interviewees indicated the need for **AI technologies should be demystified in the public imagination**. AI technologies are often simplistically referred to either as simple automated devices or as a powerful controlling and self-learning phenomenon from the near future (Alizadeh et al., 2021), but there is little understanding about how such technologies are already in use and influence different aspects of our everyday lives (e.g. HubSpot Global AI Survey, 2016).



People are constantly interacting with AI-based technologies, but they are rarely aware of this and do not always know how to distinguish AI technologies among other types of digital artefacts. While people are afraid of robots taking over humanity in the future, other types of problems of AI technologies that are manifesting themselves already go under the radar. As one of our interview partners put it: *“AI is like a magic beast – on the one hand, people have too many expectations that it is very powerful, while on the other, such already existing technologies are not taken seriously enough”*.

Part of this demystification is also the need for the end users to understand that AI systems are neither distinct entities that can act independently, nor some neutral and purely technological artefacts. There are deeper structural dynamics and power relations behind the creation of each algorithm. Some of our experts pointed out during the interviews that a successful AI literacy program for Reflective AI use should therefore not only consider the technological aspects, **but should also unveil by whom, why and with what end-goal the given algorithm has been developed**.


As already discussed, the fear of loss of human control over AI technologies has manifested itself prominently in recent years (Fast & Horvitz, 2017). Therefore, one of the biggest emerging needs that should be addressed is the question **how to ensure that end-users understand the basic principles behind AI technologies**. Furthermore, there is the need to investigate **how deep users’ understanding of such technologies should be** so that they don’t get overwhelmed by the complexity. While there is a normative consensus that end-users should be able to understand the outcomes of AI algorithms (e.g. Fjeld et al., 2020), our expert interviews suggest that it is hard to explain the outcomes and the internal logic of the algorithms in an understandable, yet not misleading or too simplistic way.

Existing approaches to making AI systems more explainable in use, while important in their own right, are not well placed to empower people to achieve a broader understanding of AI systems and the awareness of their possible effects. They largely treat this as a technical problem, or at best a problem of individual cognitive reasoning about a specific result or a given system (see e.g. Wang et al, 2019; Adadi & Berrada, 2018). They tend to neglect the role of social context in which AI is used in spite of recent studies highlighting its importance (Eslami et al., 2016; Kou & Gui, 2020). Thus, in Section 4 of this report we try to outline some more promising techniques and directions that could be better suited to address these needs.

Furthermore, one fundamental question that arose from our expert interviews is **whether or not users are really interested in learning how AI systems work**. Existing AI explainability approaches tend to underestimate the inherent effort and willingness needed by users to consciously engage into reflection on the results and the behaviour of an AI system while using it. This is in opposition to users’ expectations of a frictionless and efficient use of such systems, whose very purpose often consists in reducing cognitive complexity and helping users deal with information overload (for frictionless design see Hamilton et al., 2014; Weiser, 1994; for information overload see Koroleva et al., 2010).

To what extent people may actually consider explanations of AI systems and their results strongly depends on their willingness and ability to do so, i.e. on their ability to reflect on their use and experience of AI systems. Even when explanations are provided people may ignore them if the given results contradict their existing beliefs (Knobloch-Westerwick et al., 2020). They may still defer responsibility to an “intelligent” system as a coping mechanism for dealing with a cognitively overwhelming task or because effortless use provides an immediate gratification (Ryffel & Wirth, 2020).

Similar concerns were expressed also by the experts within our interviews. According to some of them, there is only a very small number of interested users who would want to know more about the way the algorithms work, while the vast majority of people will take the outcomes as they



are. And this is not necessarily a problem if such technologies have been checked adequately in advance. As one interview partner pointed out: “*It should be like I am on a plane. I don’t know how it works, but I feel safe, because people have checked it in advance, so I don’t need to understand how exactly it functions*”.

Others, however, see a threat in the fact that people expect digital technologies to be completely accurate and cannot adequately comprehend the idea of systems being not 100 % accurate in their estimations. This has the potential to lead to users overtrusting the results of the AI decision-making (e.g. Howards, 2020) with potential serious or even deadly consequences for them (Thornhill, 2020) as also shown in Section 2.1 of this report.

### **3.1.2 Operational principles and hidden properties of AI**

In order to demystify AI technologies and enable end-users to understand them for what they are, we recognize **the need for key AI properties to be understood by users**.

Specifically, there is a need to enable the development of appropriate mental models (Johnson-Laird, 1980) that people have of AI systems, i.e., their internal mental representations, an intuitive understanding of how the system works and behaves (Kulesza et al., 2013). Such structural mental models influence how people interpret the behavior and the results of systems they use (Normann, 1983). They guide users’ expectations, actions and behaviour based on their experience with what they consider similar systems (Normann, 1983), as well as based on social exchanges with others (Devito et al., 2018).

So far, we have identified **five key properties of AI** that need to be addressed so that people can shift their mental models about AI in a more reflective direction that better grasps the reality behind AI technologies: *sensitivity, temporal effects, non-linearity, “birds-eye-view” and privacy*.


#### ***Sensitivity***

One key challenge we see is that the fundamental principles and properties of AI – and their effects on individuals and society – cannot be directly experienced and observed in casual interaction with AI systems. For example, AI is sensitive to minor variations of input (e.g. deep learning, recommender systems), which users normally can’t observe and reason about: very small changes in training data or user interaction can cause major differences in the results (Jiawei et al., 2019). The reliability of such results thus needs to be carefully assessed, especially when they can have major consequences (e.g. health, policing) and also when they can be induced on purpose by manipulating the data in ways imperceptible to human users (e.g. adversarial attacks, see Goodfellow et al., 2017; Moosavi-Dezfooli et al., 2016; Kurakin et al., 2017; Papernot et al., 2017).

But this *sensitivity* and its consequences are not directly observable for users and are difficult to convey through isolated explanations of a given result. This induces wrong mental models with misplaced trust in results that can reinforce existing biases (Nickerson, 1998; Michael & Otterbacher, 2014) and lead to harmful decisions (Hill, 2020).

#### ***Temporal effects***

Even less observable to users are *temporal effects* of the use of AI systems. The effects of AI accrue over time and at large scale and are thus difficult to discern and understand in individual use. For example, it is difficult to observe and understand how gradually changing content recommendations over time can impact one’s beliefs and ethical judgments (e.g. becoming more polarized in online discussions or open to extremist views (Kaiser & Rauchfleisch, 2018; Ribeiro et al., 2020). Changes in preferences, perceptions of oneself and of one’s social reality that are



highly mediated by online platforms using AI, often develop at the implicit level over time and are thus difficult to consciously recognize.

### ***Non-linearity***

The related *non-linearity* of AI models is another property of AI that most people don't have a natural intuition for. Grasping the nature of exponential growth that stems from non-linear phenomena is intuitively difficult because we are not used to experiencing phenomena that change very quickly in very short time. In a similar way, it is difficult to understand that a few clicks on personal recommendations can lead to completely different content than what one would normally be exposed to or deem acceptable and get oneself quickly absorbed into (the "rabbit hole" effect (O'Callaghan et al., 2015)). This makes it even more difficult for users to develop an awareness of the need for a more conscious use of such systems or of the need for societal regulation of their design, implementation and acceptable modes of use.

### ***Birds-eye view***

In addition, in AI systems each user commonly experiences only a small portion of a system's behaviour and its results, as these are often highly dependent on personal preference profiles and users' history of interaction with the system (Hamilton et al., 2014). A "*birds-eye view*" that would make system behaviours experienced by many different users and the effects that these entail observable is not available to normal users. That makes it difficult for people to develop an awareness and understanding of how the underlying properties and behaviours of a system using AI technology may be related to harmful personal and societal effects (e.g. misinformation (Fourney et al., 2017; Allcott et al., 2019; Hassan, 2019; Fernandez & Bellogin, 2020), online radicalization (Ribeiro et al., 2020)). Thus there is little motivation and few possibilities for people to reflect on their assumptions and the behaviour of the underlying AI systems while using them.

### ***Privacy preservation***

Last but not least, a complex issue underlying all AI systems is how they deal with *privacy preservation*. The EU GDPR regulation has forced providers to disclose how a system collects, processes and uses personal data of the users, but this information and its *implications* are difficult to understand. Most critically, how AI systems can be designed and applied in *privacy preserving ways*, as alternatives to data-greedy approaches are unknown to most users. This leads to a false sense of inevitability of surrendering personal data as a trade-off for effective use - often a false dilemma resulting from biased system design choices (Larson et al., 2017).

There was a consensus between the different experts we interviewed within the *Reflective AI* project that **the level of responsibility that should be attributed towards the end-users should be limited: users could be made aware of certain issues and risks** with respect to the use of AI technologies and they need to have some **basic level of understanding of the workings of AI algorithms**. However, **structural measures (e.g. ethical guidelines, regulation) should also be put in place** that make sure that AI is developed and applied safely and responsibly by the **developers and providers of AI technologies**.

In addition to the outlined key properties and principles of AI technologies in the chapter, there is thus a need for more research to what else should end-users, on one hand, and the different societal actors such as AI designers and regulators, on the other hand, need to understand and consider in their use, design and implementation of AI systems in practice.



### 3.2 AI developers and designers

While the previous section addressed what end-users need to understand about AI in order to use such technologies in a reflective manner, we recognize that designers of AI systems should also consider what makes it difficult for people to develop this kind of understanding and capacity for Reflective AI use. We need to understand how we can design AI systems, dedicated learning environments or interventions, that enable people and provide opportunities for people to develop such kinds of reflective understanding of main AI and principles and properties.

Therefore, this section addresses the aspects that AI designers and developers need to understand about users' needs or change in their work practices to be able to support the end users better in achieving Reflective AI use.

One of the main aspects that was mentioned many times in the interviews is the fact that AI developers and designers often also **don't understand entirely how the systems they are creating make certain decisions**. With increasingly more complex algorithms used to fulfill tasks in all areas of life, the "black box" (Castelvecchi, 2016) predictive models can become so complicated that no human can understand how the input variables are jointly related to each other to reach the final output (e.g. Rudin & Radin, 2019).

This contributes to the fact that in many cases AI designers and developers can see the problems they haven't considered during the development process manifesting themselves only post-factum. Furthermore, this means that even AI designers and developers cannot always sufficiently explain a given outcome of the algorithm which makes it even harder to explain it for end-users who know almost nothing about the issue.

One of the interview partners specifically focused on UX designers who, according to him, often have very limited understanding of what AI technologies are capable of and are therefore perceiving them in a similar way as the end-users: as a sci-fi futuristic scenario and not as something that is already implemented, used and needs to be understood and explained.


This claim is supported by research that finds that UX designers struggle with both conceptual and operational knowledge of machine learning capabilities, limitations and data requirements, in order to ideate realistic applications that address end-users' needs and fit a particular context (Dove et al., 2017; Dudley & Kristensson, 2018).

The UX designer group is particularly important because they are the connection between the end-users and the AI developers and they are the ones who should link these two sides and make the technology accessible and understandable for the users. Therefore, it is crucial that designers are provided with the tools to understand how AI technologies function so that they can later create patterns or guidelines that help users to navigate the systems.

The fact that many end-users perceive AI as something hidden and magical, and take the results it provides at face value, is actually exacerbated by the currently dominant approach to user experience design in commercial practice. Driven by the necessity to increase engagement and conversions (the goal provided to by the management, see section 2.1), current **UX designs** tend to **consciously hide the complexities of the underlying system** in order to make the process as seamless as possible (Hamilton et al., 2014). Such designs nudge the customers to take the recommendations at a face value and as a result buy the recommended products without questioning the quality of the recommendation.

Including explanations into these processes is mainly done for internal purposes - e.g. for machine learning engineers, who use explainability to debug the model itself (Bhatt et al., 2020) - and not necessarily for the end users. Moreover, if one would want to include such explanations, one would need to learn how to visualize uncertainty (e.g. see Holzinger, 2018), or explain to the user that the recommendation is not 100% fitting for them, which in turn would most likely not





result in a purchase. However, hiding this information from the users violates the basic principles of UX and reflective AI design (outlined in section 4.1) such as understanding and controlling the system.

Furthermore, many interview partners mentioned the fact that AI developers **don't have the understanding or sensitivity that they are developing artefacts and technologies that can profoundly influence the individual and public life**, but rather think of their work mostly in terms of optimizing the outcomes of algorithmic processes. According to some of the experts we interviewed, the developers of AI technologies shouldn't only learn mathematical and technological operations, but should have a curriculum that also integrates philosophical, ethical and societal topics and issues for consideration. This also mirrors suggestions from recent research (e.g. Saltz et al., 2019).

The same way doctors are being trained with the idea that their work will be influencing humans and society in a dramatic way, AI developers should have a similar understanding of the importance of their role and responsibility. As one of the interviewees put it: *"We need to create a level of awareness among developers by providing them with tools to evaluate the ethical implications of their work, because so far they only want to optimize and increase the accuracy of the final results"*.

Most of the interview partners see AI developers and designers as actors with very high responsibility and ability to influence the development of Reflective AI technologies. Here they don't necessarily mean the individual designers or programmers, but rather the companies and entities that are responsible for the creation and marketing of such technologies as a whole.

Some of the interview partners suggested that the efforts towards achieving Reflective AI should start with the AI designers and developers by providing them with the right tools to understand and reflect on their own position and responsibility. Others focus more on the need for better regulatory systems and frameworks in the field of AI.


Finally, designers and developers of AI systems are often private actors and entities, even if in some regions and contexts, states and public structures are also actively participating in the development of such technologies (e.g. Europe, China). Given this, one of the biggest challenges that many of our interview partners saw in the development of Reflective AI technologies is the **tension between the private interests - namely profit maximization - and the public good**. For example, many companies need a lot of user data to make their business models work properly, thus data privacy is by logic contradictory to their own business goals and interests.

These inevitable contradictions within a market economy cannot be solved by the free market alone. Even if some interview partners suggested that increased consumer sensitivity would push the companies towards more ethical behavior and despite the attempts, especially in Europe, to create a narrative that would link the ethical behavior and the increased customer trust with higher profitability, almost all interview partners expressed the need of institutional public regulations and guidelines that would effectively control the AI development process (more in the following subchapter).

### 3.3 AI regulators

Most of the interview partners agreed that one of the very important levels of intervention in order to guarantee the development of truly Reflective AI technologies and practices, is the existence of adequate regulatory and legal frameworks. Institutions should step in, provide standards and control the development and implementation of AI technologies *before* they are made available for the end-users.

The main problem mentioned by many experts is the fact that **public regulators are very slow** and often bureaucratic, due to the nature of their work, while the technological developments are occurring at a different, faster pace. This speed discrepancy contributes to the fact the



regulations and control come in place only after severe malfunctions and problems have manifested themselves.

The need for a more democratic control over private enterprises that are developing AI technologies was formulated as follows: *“Governments should be able to access and audit the process of AI technology development. If some of the developed technologies are not benefiting or are even harming society, they should not be allowed on the market. These technologies should fulfill certain standards. It is not possible to develop technologies that are 100 % discrimination and bias free, but we should at least try [...] We should have something like the equivalent of the German TÜV [periodic vehicle control] for AI technologies”.*

However, even if the prevailing perceptions of public institutions is as slow and badly prepared to cope with the upcoming technological developments, there were experts who are closely working with the public administration in Germany, who disagree with this view and see the public administration as modern and adaptable, especially when given the right tools to deal with the emerging digitalization trends. Therefore, a productive direction of research could be to find ways to equip public servants with the knowledge and tools that would help them to understand better AI systems in order to be able to control them better.

There are different ideas about which organisations and institutions should be responsible for controlling the AI development process. While some of our interview partners point towards governments and public servants, others are looking at supra-governmental structures such as the EU or the UN. A promising development in this regard is, for instance, the recent European Commission draft legal framework on establishing trustworthy AI within the Union<sup>8</sup>.

A third group of experts addressed the need for establishing “new institutions” that are faster and better equipped for the new technological realities and that could come from civil society. However, the latter also acknowledge that civil society actors are still not well organized and the efforts there are spread across many smaller entities which makes coordinated collective actions harder. In this sense, one of the possible directions to go for would be to develop tools and formats for civil society actors to organize better together.

---

<sup>8</sup> Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence: [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_21\\_1682](https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682)





## 4. How can we design systems and solutions that support a reflective use of AI?

*The previous section has highlighted what different types of actors should be able to understand about AI in order to use it safely and responsibly, to harness its benefits and prevent harms. In this section we turn to the question: how could the design of AI systems address these needs? To this end, we propose concrete design considerations for AI systems to better support reflective use.*

How could we design AI systems to enable end-users and stakeholders to better understand AI and its consequences in order to use and govern it responsibly, harness its benefits and prevent harms? We have asked that question in a workshop to an interdisciplinary group of researchers from academia and industry; we have discussed it in expert interviews to a wider range of stakeholders from research, education, companies, media and civil society initiatives; and we have addressed it by investigating existing literature.

The insights presented in this section stem largely from the expert interviews, the interdisciplinary workshop “Reflective AI in a digital society”, written contributions from some of the workshop participants and from the subsequent analysis and ideas of the project partners. When additional observations are included based on literature (or when literature corroborates the findings from the workshop and interviews) this is supported with corresponding references.

Guiding principles for a responsible design and use of AI have increasingly been described in a rising number of documents by different types of actors (for a review see Fjeld et al., 2020). These describe high-level principles as normative requirements that AI should fulfill. Thereby a growing consensus is emerging around a set of key themes (see Fjeld et al., 2020): *privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human-control of technology, professional responsibility, promotion of human values.*

These guiding principles for responsible AI are hugely important. But it is still a challenge to break them down to operationalizable design considerations. We aimed to derive concrete design suggestions that AI systems should consider in order to implement the requirements for enabling a reflective use of AI (see Chapter 3), that we see as a “missing link” in current approaches.

Thereby, **the need to demystify AI is an overarching prerequisite for a more reflective use of AI.** As discussed in Section 3, this holds both for general perceptions of AI by laypeople as well as for misconceptions of different types of actors in using AI.

It is not only the general public that often relates AI to a “mystical” intelligence from SciFi movies, unaware that AI is present in many daily activities they perform, such as browsing on the Internet or in the feeds of their social networks. **Misconceptions about the nature and the behaviour of AI systems are also held by decision-makers when they make decisions that affect both individuals and society.**

In this section we thus discuss what should be accounted for in the design of AI systems to enable the **demystification of AI: to help users to develop a better understanding of AI systems and their actual ways of operation - and to keep control** of how their personal data are used by AI. To this end, we propose design considerations for AI systems on three main levels:

- Transparency of AI presence (“AI inside”),
- Understandability of AI (“hidden properties”),
- Control over the use of personal data in AI (“privacy preserving AI”).



The following diagram illustrates the envisioned processes of experiential learning about key hidden properties of AI described in detail in the following section. The left side of the diagram summarizes the key problems with regard to users’ perception of AI (as described in Section 3.1), while the right side of the diagram shows how these false perceptions could be challenged in order to empower end-users to use AI technologies more reflectively (as outlined in Section 4).

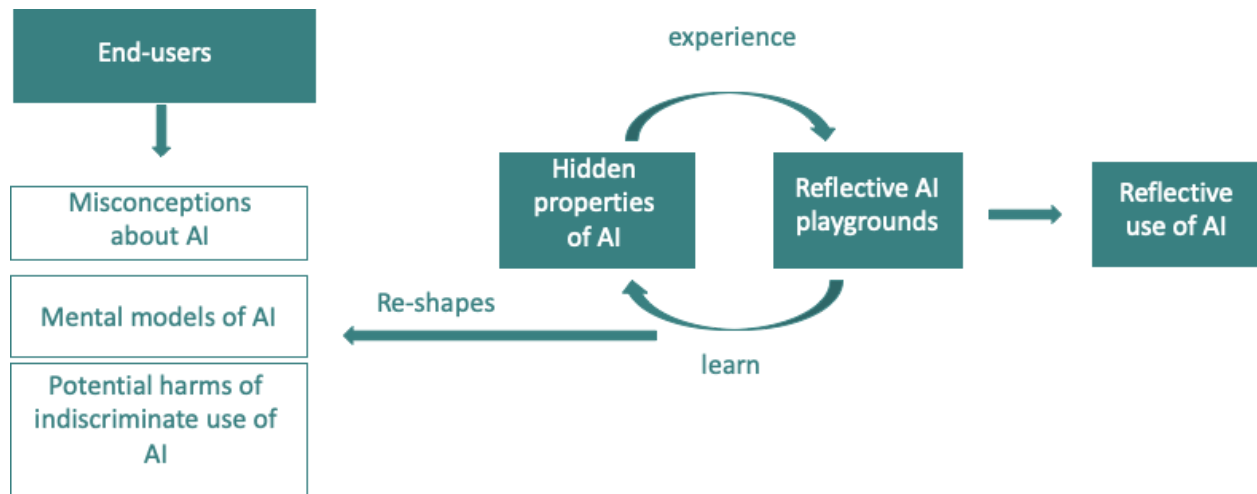


Diagram 1. Towards Reflective AI: End-users and experiential learning about hidden properties of AI.

#### 4.1. Transparency of AI presence (“AI inside”)


**Ensuring that users are aware of the presence of AI** in a system they are using is a fundamental prerequisite for demystifying AI and helping users understand its underlying nature. This is currently missing in many systems, especially those used by the general public as part of their everyday lives (e.g. Internet search, online social networks; see Section 2.2.2). A simple solution, a well-visible “label” (e.g. “AI inside”) and/or an alert signalling the presence of AI could already help as a first level of raising user attention.

An alert could notify the user if there is an AI algorithm working in the background, similar to how the GDPR requires companies to inform the user that they are collecting their data and which data is being collected. At the next level, users could be informed about the different purposes for which AI is used in the system. As one of the interview participants mentioned: *“people have a right to access this layer, trying to pull back the curtain to give an idea what is going on with their data, first step with people taking control”*.

This would ensure a basic level of **transparency of AI presence** for any given system. It could be achieved, for example, by showing an icon and then offering additional information about the underlying AI system on-demand. This is important, because if users do not know that AI is involved in the system they are using, what its capacities and limitations are, using the system unaware can lead to personal and societal harms (see Section 2.1).

At the next level of attention, the system design should make it clear and transparent to the users exactly which parts of the system functionality are based on AI and what effects this has on the system’s results and behaviour. One solution could be to provide explanatory “tours” of the system that explain its behaviour and the role of AI in it (and mandate it by regulation), similar to the guided tours of main features that are already commonly provided to new users or after system upgrades (“What’s new”) by different kinds of software .

Extending such guided tours with a particular focus on the role and purposes of the usage of AI in a given system could be done in a similar way. Another way to address this level of signaling



could be achieved by marking specific functionalities where AI plays a role (e.g. an AI icon over these functionalities) and adding short narrative explanations to them (e.g. like tool tips commonly used to explain features of existing systems).

It is however unclear to which extent users would be willing to engage with this information and how it should be presented, so that it is easily understandable for many different users. Providing this information is also likely to increase the overall information load on users, who thus might avoid considering it. These problems are similar to the provision of information about the use of personal data mandated by GDPR with explanations and settings that are difficult to understand and to use effectively (Sanchez-Rola et al., 2019; Utz et al., 2019).


How these different levels of signalling of AI presence should be best addressed, so that they actually attract user attention, motivate them to engage with the presented information, avoid information overload and make it easily understandable, are open research questions. Devising suitable solutions could build on existing research in algorithmic awareness (e.g. Alvarado & Waern, 2018; Eslami et al., 2015), human-AI interaction (e.g. Amershi et al., 2019) and persuasive communication for behavioural change (De Wit et al., 2008; Moyer-Gusé, 2008; Novak et al., 2018).

**Moreover, the awareness of the presence of AI and the purposes of its use in a given system shouldn't be seen as a sufficient goal in itself. That is only a necessary first step, a prerequisite for learning about what the system does, what for and how it uses AI and the consequences thereof.** This in turn is a prerequisite for sovereign usage and control of a system's use by the user (the principle of autonomy). And it is also a prerequisite for the users to be able to critically assess and challenge system results and provide feedback to system developers and providers.

To implement this approach it's not only the challenges of understandability, user engagement and information overload that need to be resolved. Whether the described kinds of information will be willingly provided by the companies to the users is not entirely evident and companies might not be motivated to do so. Revealing this information should be in the interest of the companies themselves as it can increase users' trust in the AI system and its results, as well as in the company itself. But as some interview participants described, many companies are "opaque and secretive" and their services are designed in a way that the users should not be aware or informed of what is happening in the background. So this kind of transparency would likely need to be mandated by regulation.

The change in how AI systems are perceived by people is a profound challenge - it requires a fundamental shift in the minds of users as well as in the attitudes of the companies. **Transparency of AI** calls for revealing **what sort of technology is being used in a specific case, how it is used to benefit the individual and what the risks of this technology are.**

Moreover, as highlighted in the interviews, a *steep learning curve* in understanding AI is expected: "Once you have seen the explanations a few times, you don't need them. When you have a new customer, you can explain, but after a certain point, maybe they have gained trust in the system, and don't need explanations any longer" (as an interview participant put it). This suggests that the **explanations about the presence and purposes of AI** provided at this first level of awareness likely need to be scaffolded (Quintana et al., 2004; Jackson et al., 1998; Sharma & Hannafin, 2007) at different levels of complexity. Rather than aiming at providing a full-sized understanding all at once, they could lead the users to successively better understanding of what they need to be aware of and understand in order to use the system competently, safely and responsibly.



This raises the question of what kind of information and what kind of explanations could (and should) be provided for this purpose. Moreover, in Section 3 we laid out why an understanding of the implications of the use of AI in a system requires people to understand the underlying principles and properties of AI that are normally hidden.

This leads to the following questions: **How could the underlying operational principles and normally hidden properties be exposed and made understandable to the users?** How could this be achieved so that users internalize this understanding in new, **more appropriate mental models of what AI is, how it operates and what benefits and risks it carries?**

## 4.2 Understandability of operational principles, properties and risks of AI

Once the users are aware that there is an AI algorithm working in the background and for what purposes it is used, they would need to be explained what the AI system does, how it does it and which risks this may possess. This is the next level of user awareness of AI. How much and which parts of the system to explain to the users, is still a question to be answered. It is not enough to just inform people about the consequences of an unreflective use of systems employing AI (e.g. the risks of overtrusting the system results when taking decisions, the potential effects on one's beliefs and perceptions; see Section 3).


If people are presented with information that contradicts their existing beliefs and opinions, they are likely to refute it (Nyhan & Reffer, 2010), as opposed to information that confirms what they already believe in (the so-called confirmation bias). Similarly, as research in persuasive communication has shown, a number of factors beyond the information content influence the extent to which a given message (information) is ultimately accepted by a person (Naul & Liu, 2019). At the same time, narrative and entertainment education strategies can be a promising approach for persuasive communication, if their design appropriately considers specific factors that influence the likeliness of acceptance by the users (Moyer-Gusé, 2008; Slater & Rouner, 2002).

### 4.2.1 Explaining operational principles

In order to make the explanations of the risks and potential harms credible and comprehensible to users, we have argued that it is essential that they also develop some level of understanding of how the underlying AI algorithms actually work (Section 3.1) - let it only be in terms adapted for laypeople. The mathematical principles and intricacies of AI algorithms can be difficult to understand even for experts. But the main operational principles of many AI algorithms, their conceptual logic, could be explained in terms suitable for laypeople without delving into the mathematics behind them.

Devising such narrative explanations in ways that are understandable for laypeople but true to the underlying operational principles of an AI algorithm is however all but trivial. For example, in order to explain how a collaborative filtering algorithm works on a recommendation website, one could explain the underlying conceptual idea of item-based recommendation in relatively simple terms, as one of the interview participants mentioned: *"We just count what you have been buying before, compare it to other people and show it to you"*.

But while this kind of explanation of a specific recommender technique is simply understandable and doesn't overwhelm the user, it also carries the risk of oversimplification. If that's all there is to it, what's there to worry about? How can the risks associated with unreflective design and use of recommender systems be then motivated and made comprehensible to the users (e.g. the problem of clickbaiting, or the risk of radicalization on YouTube)?



This illustrates a major challenge: **How to devise explanations of operational principles of AI that are comprehensible for a wide-range of users, while sufficiently precise to set the ground for understanding subsequent explanations of potential risks?**

One way could be to **start from explanations of the specific results and system behaviour that the user can observe and expand these with narratives about their possible causes and consequences.** Using **metaphors and visualizations** to communicate these (e.g. Segel & Heer, 2010) could also help to make it easier for people to connect to existing concepts that they are familiar with. This could also make users more motivated to explore and learn about the system behaviour more closely, as opposed to getting them scared off by complex (often mathematical) concepts that are usually part of AI algorithms.

But to provide such explanations that make the workings and consequences of AI systems understandable to lay end-users and stakeholders, AI models need to be interpretable by design. Research on explainable AI has given a lot of attention to finding ways to explain the results of machine learning models that are normally opaque and difficult to interpret (“black boxes”). But such post-hoc explanations of black box machine learning models are often unreliable and can be misleading even for AI experts (Rudin, 2019; Rudin & Radin, 2019).

Research on interpretable machine learning has a long tradition, often under different names (e.g. Holte, 1993; Freitas, 2014) that is easily overlooked in current developments. Recent approaches such as representational learning have also shown how existing machine learning techniques that are not interpretable (e.g. deep learning) could be re-conceived in ways that provide interpretability by design (e.g. Wang & Rudin, 2015; Zhang et al., 2018). Such approaches are of crucial importance for enabling a reflective use of AI, because interpretability is not only a prerequisite for enabling end-user understanding. Ensuring interpretability by design is also required for showing how the internal workings of AI models relate to both expected benefits and potential risks. Uncovering and making such relationships observable is crucial for enabling critical reflection.

An important aspect here is also to **show not only the possible risks, but also the benefits of using AI-based systems.** As an interview participant put it: *“For example, Youtube is dangerous, you can get radicalized due to recommendations that show you more and more of the same stuff, but empowering too, as you get education on a lot of stuff, very liberating, this could be something you could leverage and try to bring people to be more interested in what is happening, by saying what is good about it.”*

A certain level of **adaptability to the needs and capabilities of different users** could also be provided with different levels of detail of explanations to choose from (e.g. mathematical details on-demand). This would also align well with the scaffolding principle: allowing users to choose different levels of difficulty or complexity of explanations as they gain more experience with the system, as that has worked well in other domains (e.g. computer-supported learning (Jackson et al., 1998; Sharma & Hannafin, 2007; Quintana et al., 2004)).

No matter how detailed, the **explanations of AI behaviour should be relatable to the user,** to their current experience and current context. If the users can recognize how the explanation actually refers to the results that they were shown (e.g. recommendations received) or the data they provided, then the consequences and the workings of the underlying AI system are likely to be grasped more easily and more willingly. Moreover, constructivist theories of learning (Ackermann, 1996) suggest that explanations should be **interactive** and that users should be able to have hands-on experience with the systems. Interactive recommender systems (He et al., 2016; Jugovac & Jannach, 2017) and interactive machine learning (Dudley & Kristensson, 2018) have shown to provide important benefits in users’ understanding of AI technologies.

Explainability and interactivity go hand in hand, as interacting with an AI system will provide more insights into its inner workings. Interactivity in such a way also benefits user trust and acceptance (Schnabel et al., 2020). For example, users could explore what happens in the system if they change some of its parameters. This could help to transfer the abstract concepts to actual use cases as well as to increase the motivation of the user to explore the workings of the algorithm. Actual learning from experience happens after people reflect on what they have had experience with (Kolb, 1984). After interactively engaging with the system, users would not only understand it better, but also be better able to consciously decide if they are willing to use the system at all. As one of the participants mentioned: *“In our data relation platform<sup>9</sup>, we show the user before they donate their data what this data is about, we visualize it and let the user interactively explore, before they decide if they want to donate his data, or not.”*

Finally, as shown in Chapter 3.2, the complexity and uncertainty of AI results is often hidden in order to simplify and make the results more easily accessible and usable for the users (e.g. using recommendations to ensure conversions from visitors into paying customers). However, such practices go against the principles of Reflective AI design that requires users to understand and be in control of the technology they are using. Therefore, there is also an emerging need to develop ways to make AI developers and UX designers aware of what the users actually experience when they see the results of AI algorithms.

Accordingly, the user experience pipeline would benefit from being entirely rethought, so that it not only explains in an easy and interactive manner what the system does, but also does not result in overloading the users (Koroleva et al., 2010) which could refrain them from fulfilling their goal (e.g. choosing and buying a suitable product). This is an important concern both for the users themselves and for the companies that employ such AI systems.

Rather than considering AI transparency and explanations as an add-on, by rethinking the entire user experience of AI systems, designers could develop novel ways to ensure explainability without overloading the users. As one interview partner put it: *“You can have a box with a dry explanation, but the alternative is in the interface of the system, designers are so innovative in showing content, so they can develop a solution which is interactive”*. User experience designers could create **new design patterns to visualize and reflect uncertainty**, which is pertinent to results of any AI system, in a way that users understanding this information, can still make their own decisions.

#### 4.2.2 Enabling users to learn about key properties of AI

In order for users to really grasp why and how AI systems can lead to specific risks and harms they need to develop an understanding of key properties of AI that are normally hidden from users. As summarized in Table 1, these include: the *sensitivity of AI algorithms*, *non-linearity* and *temporal effects*, what we term the *“birds-eye view”* and the *privacy preservation* (see 3.1.2).

| Key hidden properties of AI users should understand |   |
|---|---|
| Sensitivity   | AI techniques, e.g. deep learning (LeCun et al., 2015), recommenders (Jannach et al., 2010), are highly sensitive: very small changes in training data or user interaction can cause major differences in the results (Jiawei, 2019). Sensitivity can have serious consequences not only in commonly assumed cases (e.g. health, policing), but also broadly (Liu et al., 2019). <b><i>By helping users become aware of sensitivity we can correct mental models and avoid misplaced trust in results that can reinforce existing biases (Nickerson, 1998; Michael &amp; Otterbacher, 2014) and lead to harmful decisions (Hill, 2020).</i></b> |

<sup>9</sup> This refers to the DataSkop project of AlgorithmWatch: <https://algorithmwatch.org/en/dataskop/>






|                                    |  |
|------------------------------------|--|
| <p><b>Temporal effects</b></p>     | <p>Effects of AI techniques accrue over time and at large scale and are thus difficult to discern and understand in individual use. For example, it is difficult to observe and understand how gradually changing content recommendations over time can impact one’s beliefs and ethical judgments (e.g., causing polarization in online discussions or openness to extremist views (Kaiser &amp; Rauchfleisch, 2018)). <b>Allowing users to experience time-lapse versions of AI could help them reflect on the dangers of temporal effects and the related non-linearity of AI (e.g. the “rabbit hole” (O’Callaghan et al., 2015)), leading to implicit changes in perceptions of social reality.</b></p>  |
| <p><b>Non-linearity</b></p>        | <p>Grasping the nature of exponential growth that stems from non-linear phenomena is intuitively difficult because we are not used to experiencing phenomena that change very quickly in very short time. In a similar way, <b>it is difficult to understand that a few clicks on personal recommendations can lead to completely different content than what one would normally be exposed to or deem acceptable and get oneself quickly absorbed into</b> (the “rabbit hole” effect (O’Callaghan et al., 2015)). This makes it even more difficult for users to develop an awareness of the need for a more conscious use of such systems or of the need for societal regulation of their design, implementation and acceptable modes of use.</p>                              |
| <p><b>Birds-eye view</b></p>       | <p>AI techniques have effects that are visible only from a birds-eye view. Each user experiences only a small portion of a system’s behaviour and its results, as these are often highly dependent on personal preference profiles and history of interaction with the system (Hamilton et al., 2014). That makes it difficult for people to develop an awareness and understanding of how a system using AI may be related to harmful personal and societal effects (e.g. misinformation, online radicalization (Ribeiro et al., 2020)). <b>By offering the bird’s-eye view, we could allow users to become aware of their overall impact on issues such as misinformation and online radicalization (ibid.)</b></p>  |
| <p><b>Privacy preservation</b></p> | <p>AI techniques can be designed to protect user privacy but these possibilities are largely unknown to users. This allows companies to present the need to surrender personal data in return for effective use of an AI system as an inevitable necessity. The EU GDPR legislation has forced providers to disclose how a system collects, processes and uses personal data, but its implications are difficult to understand and their use by AI is not specifically described. <b>By providing users with insights into the workings of privacy-preserving AI they could learn to reflect on the necessity of surrendering personal data in return for system effectiveness, often a false dilemma resulting from biased system design choices (Larson et al., 2017).</b></p> |

Table 1. Key hidden properties of AI that users need to understand in order to use AI reflectively (see Section 3.1.2 for motivation and details).

**But what could be done to enable users to grasp the nature of such properties of AI and their implications at the personal and societal level?** We believe that this can be only partially addressed within the design of AI systems themselves and exposed to users during normal use.

Grasping and learning about these issues **requires willingness and effort to consciously engage into reflection** about the behaviour of an AI system *while using it*. This is in opposition to users’ expectations of a frictionless and efficient use of such systems, whose very purpose often consists in reducing cognitive complexity and information overload. This doesn’t mean that the system design couldn’t consider such aspects at all (see recommendations in the previous section and an example at the end of this section).



But it is unlikely that people will provide the attention and effort needed to correct their mental models based on recognizing and understanding the hidden properties of AI and their effects and consequences, during actual use of complex AI systems. Reflection commonly occurs when there is a “breakdown” in one's experience, a problem or an inconsistency that cannot be resolved within one's existing frame of reference (see review in Baumer, 2015). Preventing such situations from occurring is the very goal of system design (seamless design), understandably so.

Thus, creating effective triggers for reflection during the use of an AI system is likely to be difficult, since both users and system designers tend to generally share a common goal: an easy, effective and enjoyable use - that avoids inconsistencies and conceptual “breakdowns”. This is also where we see a critical limitation of current approaches to explanations of AI systems and their results.

Below, we present two different approaches to how this could be addressed. One is based on the idea of a separate learning environment for experiential learning about AI. The other discusses how specific hidden properties could be made more transparent and observable during the use of a given AI system, on the example of news recommenders.

### Example approach: Experiential learning environments for Reflective AI


**We propose that dedicated interactive learning environments are needed that allow people to experience and reflect on the key properties of AI systems and their possible effects on individuals and society.** They should stimulate people to reflect on these experiences and develop new *mental models of AI* - i.e. engage them in *experiential learning* (Kolb et al., 1984; Morris et al., 2019). Developing such mental models, overall ideas of how AI systems behave and how they can lead to negative personal and societal impacts, would allow people to more competently and reflectively use AI systems in everyday life, to harness AI benefits and avoid harms.

The development of a mental model is a highly experiential process in which mental shortcuts and approximation rules are formed that allow people to deal with new, unfamiliar situations by relating and comparing them to similar experiences and their conceptual models thereof that have developed over time (Johnson-Laird, 1980, 1983; Norman, 1983; Kulesza et al., 2013). This may also explain why explanatory approaches to ‘teaching’ the general public about AI are not so successful; people may not only lack the capacity or willingness to learn about AI systems, but a **pure information-based approach does not allow for experiential learning, i.e. learning through experiences and reflection upon them.**

An **environment for experiential learning about AI** should reproduce the behaviours of different AI techniques regarding the key hidden properties of AI such as *sensitivity*, *temporal effects*, *non-linearity*, *the birds-eye view* and *privacy preservation* - in situations representing real-world contexts of use. It should allow users to interactively explore how the behaviour of the system changes depending on their actions and the changes in main parameters influencing its behaviour. And it should allow users to discover how due to such properties an unreflected use of AI can lead to personal and societal harms (e.g. misplaced trust, radicalization, misinformation).

For example, for experiencing *sensitivity*, such a learning environment could allow users to explore how very small changes in input can lead to big changes in results. For *non-linearity*, how small changes in one's actions (e.g., viewing specific videos, following specific users) can create big changes in recommendations. For *temporal effects*, it could enable users to observe how system use over time could influence perceptions of oneself or impact their attitudes to specific content. For *birds-eye view*, it could provide simulations of results that other users





would see based on different interaction paths which could be explored by the users. For **privacy preservation**, it could allow users to experience the results of the system with and without privacy preservation, based on their choices which data should or should not be processed.

In line with the processes of experiential learning (Kolb, 1984), **being able to personally experience and observe the properties and behaviour of different AI techniques** (e.g. recommender systems, image recognition) **in such a way would enable people to reflect on and re-construct their mental models of AI systems**. It would allow them to reflect on their assumptions and misconceptions regarding their functioning (e.g. deterministic vs. probabilistic nature) and to develop an understanding of the underlying nature of the results such systems produce (e.g. factors influencing result sensitivity).

Such reflection would lead to changes in users' conceptualisations thus resulting in mental models that are better aligned with the actual behaviour of AI systems and in an informed awareness of possible effects of their indiscriminate use. This could help people construct more accurate mental models of AI systems, thus making them more apt to appropriately deal with AI systems and their results in their professional and private life.


For example, investigators using a facial recognition system could become more cautious in reaching conclusions on potential suspects based on the system output by considering the quality of the input image and the situation in which it was taken or the differences in reported confidence levels between different results. Viewers of YouTube videos could become more consciously selective when choosing which of the recommended videos to click and develop an understanding about what type of content they tend to approve of and why.

Such a kind of environments that enable and stimulate experiential learning about AI systems we thus term "**Reflective AI playgrounds**". The notion of a "reflective playground" embodies several key concepts that are crucial to our approach and differentiate it from related work.

Much like the provision of explanations in AI systems doesn't mean that users will actually consider them (e.g. if contrary to personal biases (Knobloch-Westervick et al., 2020), so do the envisaged playgrounds need to motivate people to use them and learn by reflecting on their experience within them. While reflection is commonly considered to be triggered by a negative experience of encountering a problem (a "breakdown" (Baumer, 2015)), e.g. in one's use of a system and an incongruent experience thereof, building on playful curiosity could be a more fruitful strategy for raising user's interest in exploring and re-examining their understanding of AI systems and their consequences.

The notion of a playground refers on one hand to the idea of inviting the users to a playful exploration of the presented environment. It builds on game-like elements and strategies that address positively connotated motivations (e.g. discovery, play, achievements, puzzle solving, helping or socially connecting with others). Game-like elements have been successfully applied in non-game contexts to stimulate motivation and engagement in so-called gamification and serious games in many domains (Hamari & Koivisto, 2019; Böckle et al., 2017; Koroleva & Novak, 2020).

Persuasive systems and serious games research have shown that strategies that promote immersion and self-affirmation increase self-motivated learning (Baptista & Oliveira, 2019; Naul & Liu, 2019; van Koningsbruggen & Das, 2009). Entertainment education strategies are generally more effective than information-based strategies, especially if target audiences are not naturally interested in a topic (Moyer-Gusé, 2008). Devising effective prompts for reflection can build on experiences from persuasive communication (De Vit et al., 2008), visualisation (Novak et al.,



2014) and the design of interactive systems for stimulating behavioural change (Novak et al., 2018; Koroleva et al., 2019; Böckle et al., 2018).

Similarly, much as playgrounds in the real-world are places of social activity, so has social interaction and exchange been highlighted as an important facilitator of both experiential learning and reflection (Obrenović, 2012; Ploderer et al., 2014; Novak & Peranovic, 2004). The crucial role of social context and collective activity has also been stressed in a recent study of how users as a collective make sense of AI systems in their own community (Kou & Gui, 2020). In fact, important large-scale AI systems are deployed and/or used within online communities and social networks (e.g. YouTube recommendations, Facebook post filtering).

Constructivist approaches to learning have demonstrated how people learn and construct mental models of the world around them through creative experimentation, co-designing and sharing (Ackermann, 1996; Resnick et al., 2000). Accordingly, playgrounds for experiential learning should be conceptualized as social environments that not only involve users in playful learning with and about AI systems as individuals, but enable them to discover, share and discuss their observations with other users and researchers.


Such Reflective AI playgrounds would enable people to experience the hidden principles and properties of AI and understand how they contribute to negative personal and societal effects. This would contribute to a more responsible societal uptake and beneficial use of AI. They could be extended by researchers to cover a variety of AI cases. They could be provided as a learning resource for students of all disciplines and offered as a training module for employees of organizations using AI. Policy makers could mandate their use to support a responsible use of AI (e.g., requiring providers to offer such playgrounds as a “training” space for users). Ultimately, this could help people to deal with online manipulation and misinformation, and become more empowered to participate in democratic processes, including the debates about AI regulation.

### Example approach: Design issues for Reflective AI in recommender systems

Another approach to help users learn about the hidden properties of AI is to consider how the effects of specific hidden properties of AI could be made more transparent through changes in the design of AI systems themselves. **A case in point is the design of recommender systems for news recommendations with respect to personalization and diversity.**

Many AI-driven recommender systems in the field of news recommendation optimize for engagement and employ collaborative filtering (Bernstein et al., 2020). Consequently, normative considerations with respect to diversity in *sources* and – maybe even more importantly – *perspectives* are missing. Personalizing a recommendation is a way for the companies to make sure that the user is more likely to buy a certain product, or likely to read more articles in a newspaper recommendation service. However, as the users are likely to consume more of the same type of product or information, they are likely to get a narrow view on the topic or product category, although there are many more options available, which might lead to adverse consequences described in Section 2.1.

It is important that every citizen has access to a wide range of news sources and perspectives. AI-driven algorithmic news recommendation could form a risk to a well-functioning public sphere, if it leads to a significant reduction in the diversity of news a citizen is exposed to. Concretely, if algorithmic curation leads to a situation in which users are only confronted with a perpetual echo of their own thoughts and beliefs, the so-called filter bubble (Pariser, 2012), important values such as societal cohesion and tolerance are at stake.



When browsing information, users are often not aware that the same website can look totally different for a different kind of user (the **lack of the birds-eye view**), and simply consume the information that is offered. Therefore, in addition to the transparency of the underlying system described in the previous sections, **there is a need for transparency regarding the positioning of the recommendation with regards to their whole spectrum**, so that the user can have a broader spectrum of options and choose a different alternative if needed.

A system should inform the users where they stand with regards to others, similar to how a user knows in which part of the website he or she is (e.g. by using the breadcrumbs or the navigation map). Additionally, the algorithms could also be tailored to show the opposite alternatives, things that the user might not like in the first place, but to inform that other opinions and options still exist. For example, one participant from an organization that moderates hateful speech online mentioned how people *“react strongly when they are confronted with a different view, but in some cases there is still space for the person to see a different reality”*.

To integrate such a transparent view and more diverse recommendations into existing systems several challenges need to be solved. First, in order to show to the user his position with respect to others, the whole spectrum needs to be defined, which for some contexts, such as political views, could be a very contested endeavour. Integrating the normative considerations is also challenging, because measuring and optimizing perspectives in news coverage is very difficult to implement at scale (Vrijenhoek et al., 2020).


Measures of diversity can for example include representation of minority actors featured in the news article, diversity in news frames, or a balance between opinion pieces and factual news stories. Developing diversity-optimizing news recommender systems comes with the risk of poor performance or becoming too paternalistic. It is thus necessary to develop novel metrics that can be combined with extant measures of user engagement and user satisfaction. Transparent and responsive user-interfaces are also of crucial importance to ensure that users accept and value a diversity-optimizing news recommendation system.

Second, revealing a more diverse recommendation set to a person might be a double-edged sword. Research has shown that there are several types of reactions when people understand that the information was tailored to them: some don't care, some don't want it, and others feel that the recommendation is not targeted enough. This contributes to an interesting trade-off *“on the one hand, people think that the recommendations are spooky, and on the other hand, they think they are not good enough”*.

This trade-off is further complicated by the fact that people don't like to think that their actions are predictable, and that they received the same recommendation as many others, causing such strong emotional reactions towards personalization. Therefore, it needs to be researched and defined how to inform people that they get personalized recommendations, but in a mindful, careful way. Here, possible solutions could include providing **interactive tools** which would visualize a search history of a person and the recommendations that person would receive, but also allow the user to change the history to completely different content and to observe the impact of the change on the recommendations.

### 4.3 Control over the use of personal data in AI (“privacy preserving AI”)

The need to give users control over the use of their personal data in AI and to educate them about the possibilities of privacy-preserving AI is crucial for ensuring that the guiding principles of autonomy and human control over technology can be fulfilled. Therefore, although this aspect has been already mentioned in the previous section on hidden properties of AI it merits a closer look.



Today many online platforms offer no options to disable the personal data collection. GDPR in Europe is a big positive step towards privacy protection, but most practical applications are difficult to understand and not user-friendly. As a result, too often users still give consent to personal data collection unwillingly just because they want to use a particular service and feel they have no choice than opt-in or not use it (Habib et al., 2020). Allowing users to **effectively control** whether and to what extent to contribute or allow access to personal data is of utmost importance. Not only is it a foundation for user trust, it is also a prerequisite for building an understanding of the underlying workings of the system and the consequences of its use.

But the opt-in principle and the configurability of permissions to access specific types of personal data are only a first step. **Real user control can only occur if the system has adequately explained its workings to the user, the purposes of using personal data by AI - and the benefits and consequences of this use.** While this holds for all types of systems in general, it is especially important for AI systems (see 4.1-4.2.2).


Additionally, transparency regarding the possible actions of the user should be provided. For example, if the users perceive a system as not being fair in the treatment of their data, they should be informed what options they have, apart from not using the system at all. Ideally, users should be provided with possible steps they can take to protect their data or at least report their concerns to the system owners and regulators. Such options need to be **effectively actionable**, i.e. they need to allow users to effectively exercise them without being overwhelmed by their complexity.

Understanding the consequences of one's actions is also critical, as an action might cause irreversible consequences such as not being presented with the same information anymore: *"If I say that I do not like this artist, I think I will not see this artist ever again. And that is drastic."* Similarly, rather than complex bureaucratic texts, showing concrete examples of the effects of specific privacy choices for the system results and behaviour would make it much easier for users to understand the stakes involved in a given case and make informed choices.

User control can also provide for an important channel of communication between users and the designers of AI systems. **Studies show that users prefer to be able to decide and modify how an AI system works (e.g. changing the recommendation strategy of a recommender system) and how their personal data is used/shared** (Mohallick et al., 2018; Su et al., 2016). Instead of offering fully automated systems, incorporating users more in the decision making process and being transparent to the users about the data collection and usage, has positive effects on users which should be in the very interest of companies using AI in their systems.

AI systems that give more control to users may also help to decrease their privacy concerns and increase the trust in the system and its service providers (Mohallick et al., 2018). Giving the users the option to be involved in the decision making process or to modify the system properties is also an important aspect to consider for learning. Moreover, by having a *"human in the loop"*, performance of AI technology can be improved as humans and AI have different qualities in detecting and fixing prediction errors. This means that AI technologies should support efficient correction, learn from user behavior and update and adapt cautiously (Amershi et al., 2019).

Most users but also many companies are **unaware that privacy-preserving AI techniques exist** that can protect personal data while allowing AI applications that require them to safely and securely process them. This leads to the false dilemma that taking advantage of AI benefits must come at the expense of privacy and associated risks.



The application of privacy-preserving techniques in AI (e.g. application of homomorphic encryption (Bonawitz et al., 2017), differential privacy (Dwork, 2008), secure multiparty computation (Lindell, 2020) and federated learning (Bonawitz et al., 2019) has already been successfully demonstrated for a range of AI methods (Aslett et al., 2015; Hesamifard et al., 2017; Hesamifard et al., 2018; Gilard-Baachrach et al., 2016) and use cases where sensitive data needs to be processed but protected (e.g. Jagadeesh et al., 2017; Mohassel & Zhang, 2017). Solutions have also been demonstrated that don't sacrifice accuracy for preserving privacy (Wang et al., 2015) as well as approaches that protect privacy by minimizing data requirements in the first place (Larson et al., 2017; Chow et al., 2013).

Privacy-preserving AI techniques carry great promise for harnessing AI benefits and preventing potential harms, but they yet need to become a norm rather than an exception both in AI research and practice. **Educating companies, researchers, general users, decision makers and policy makers alike, about the possibilities of privacy-preserving AI and the principles of their operation could dramatically shift the wrong perception that surrendering privacy is a necessary sacrifice for taking advantage of AI benefits.**

This could lead to both a better uptake of privacy-preserving AI in practice, to increased trust in AI systems that use it, as well as to better regulatory solutions. How this education and awareness could best be achieved is an open question.

Privacy-preserving techniques are technically complex and difficult to understand even for experts. **How the underlying principles of such privacy-preserving techniques and their implications in practice could be explained to a wide-range of users and stakeholders with and without technical background is an open challenge.** It is a difficult but an extremely important challenge that should be taken up by research. Helping users, AI developers, system providers and regulators understand the principles and possibilities of privacy-preserving AI could go a long way to help overcome the current binary choice of “*opt-in or don't use it*” that users unwillingly face in many AI applications.

AI research has also demonstrated approaches that allow end-users themselves to protect their privacy by altering data in ways which do not decrease its value for AI applications, but introduce privacy protection for the personal data they contain (Choi et al., 2017).

Moreover, **providing AI solutions that implement privacy-by-design and minimize personal data requirements is also in the best “pragmatic” interests of companies that provide AI services, because that reduces risks and liabilities associated with data security** (Larson et al., 2017; Chow et al., 2013). This suggests that rather than viewing privacy and AI as a dichotomy, future research should ask: **How can we design solutions that protect individuals, but still allow companies, governments and society to harness AI benefits?**

## 5. Work practices in AI design, organisational and structural changes

While the previous section dealt with concrete principles and recommendations for Reflective AI design, this chapter takes a look at the broader organisational, institutional and structural changes that need to happen to ensure the development and deployment of Reflective AI technologies.

First, we take a look at how **designers and developers can improve and create new work practices** so that the AI systems they design can better fulfill the described design requirements for Reflective AI. Furthermore, we consider the **organisational changes** that would need to occur within companies and other organisational actors that develop AI technologies.

Finally, we describe the broad **structural and institutional changes** needed for the establishment of Reflective AI technologies and practices. As in the previous chapter, the inputs here are largely generated through expert interviews within the Reflective AI project or through written contributions from the participants in our workshops.

### 5.1 (New) work practices of AI designers and developers

In addition to and in accordance with the Reflective AI design principles outlined in section 4.1, we believe that AI designers, on the one hand, and AI developers, on the other, should improve their existing work practices. We have identified the following improvements that could help both a reflective use and design of AI and that will be elaborated further in the next chapters:

- 1) Supporting user experience designers in learning about AI
- 2) Integration of ethical awareness into AI development and teaching
- 3) Integrating interdisciplinary approaches to consider context of use in AI design

The following diagram summarizes the main problems that AI designers and developers face when creating new AI technologies (outlined in Sections 3.2). It also illustrates the possible solutions in terms of work practices (circled in green) that are discussed in the following sections.

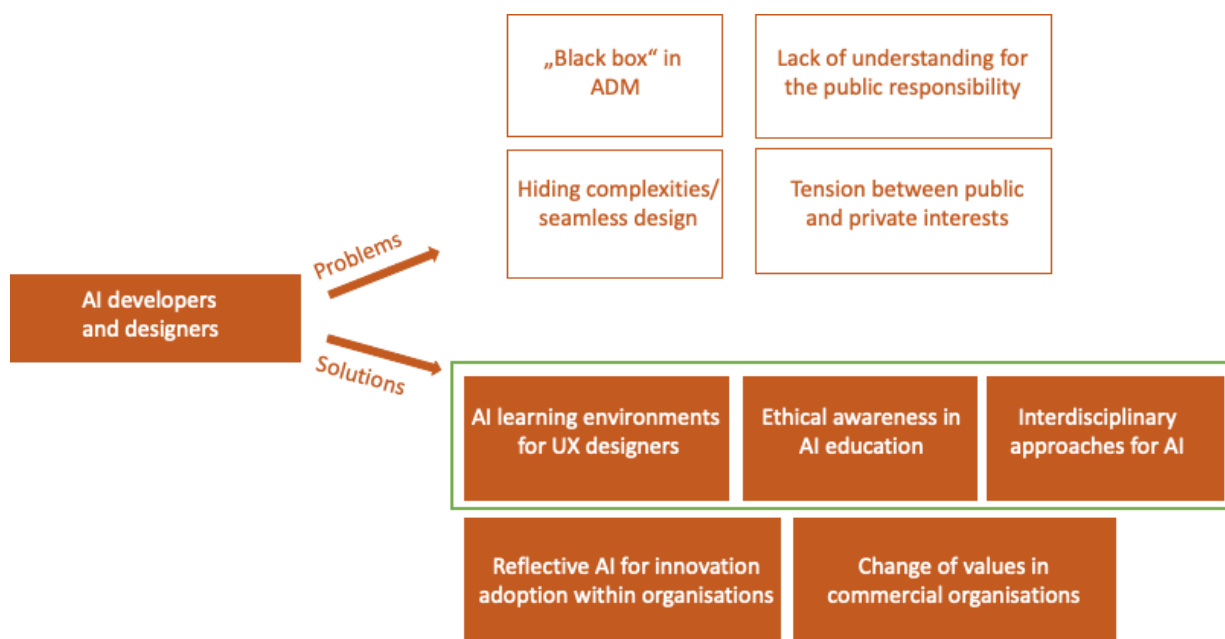


Diagram 2. Towards Reflective AI: Problems and solution approaches regarding AI developers and designers in terms of (new) work practices.





### 5.1.1 Supporting user experience designers in learning about AI

As pointed out in section 3.2, one of the main challenges for user experience designers is that they themselves do not always know or fully understand how the AI algorithms work. Furthermore, often there is no closer collaboration between them and the developers of the AI systems. In fact, the explainable user experience interface can only be developed in a close collaboration between the user experience designers who are skilled in presenting information to the end-user and the system developers who include the explainability as one of the goals when they design their systems. Therefore, ensuring that the designers understand the systems better as well as work closely with the AI developers, is another fundamental shift to the current state of things.

In order to achieve this, in line with constructivist learning theory, one of our interview partners suggested the idea of an interactive environment where the designers can learn about AI in an experiential scenario. Only if the designers understand the basic principles of AI themselves (e.g. as outlined in 3.1.2), will they be able to develop the necessary new design patterns to ensure explainability and transparency of the system for the end-users (see section 4.2.2.1 on the need of new design patterns). As shown and discussed by Winter and Jackson (2020), approaches helping designers to develop their knowledge skills through active experimentation with machine learning techniques seem a promising way forward in this regard. These experiential learning approaches and interactive environments could be furthermore created in a way to encourage and foster the direct exchange between system developers and AI designers, giving the latter the opportunity to provide feedback and requests for system improvements. A similar setting has already been implemented and tested by one of our interview partners: *“I do workshops with designers....they play around with things and see what they can do and not, then they come with recommendations of how they can change things”*.

### 5.1.2 Integration of ethical awareness into AI development and teaching

As shown in section 3.2, one of the main problems with the current development of AI techniques and technologies is that the developers mostly aim at increasing accuracy, but often neglect the ethical considerations about the outcomes of their algorithms. Such tendencies increase the risk of developing algorithms that have harmful (unintended) effects for individuals and society as a whole (as demonstrated in section 2.1). To counter this, developers should be, on the one hand, aware of the existence of such ethical risks and discussions. On the other hand, they should be required to evaluate the ethics, possible biases in the data sets that they use to train the algorithms and overall implications of their work with appropriate methods and tools.

One important way to achieve the awareness needed among the AI developers community is by integrating ethics in the machine learning courses and curriculums. Currently, this is not the standard for the vast majority of such courses. A study by Saltz et al. (2019) analyzing the machine learning and data science courses in top U.S. universities found that only about 20% of them integrate ethical aspects. In the same study, after conducting a systematic literature review, the authors identified 10 key ethical questions that could help AI developers contemplate ethical situations and tested them with a pilot of 85 students. The students were able to better identify ethical dilemmas in the machine learning sphere by using these guiding questions when approaching new assignments. This suggests that integrating these or similar ethical questions and considerations could provide useful guidance for developers both during their education, but also within an organisational setting.

| Challenge                    | Theme                               | Questions   |
|------------------------------|-------------------------------------|---|
| Oversight related challenges | Accountability & Responsibility     | 1. Which laws and regulations might be applicable to this project?  |
|                              |                                     | 2. How is ethical accountability being achieved?  |
| Data Related Challenges      | Data Privacy and Anonymity          | 3. How might the legal rights of organizations and individuals be impinged by our use of the data?              |
|                              |                                     | 4. How might an individuals' privacy and anonymity be impinged via aggregation and linking of the data?         |
|                              | Data Availability and Validity      | 5. How do you know the data is ethically available for its intended use?  |
|                              |                                     | 6. How do you know the data valid for its intended use?   |
| Model Related Challenges     | Model and Modeler Bias              | 7. How have you identified and minimized any bias in the data or the model?                                     |
|                              |                                     | 8. How was any potential modeler bias identified, and then if appropriate, mitigated?                           |
|                              | Model Transparency & Interpretation | 9. How transparent does the model need to be and how is that transparency achieved?                             |
|                              |                                     | 10. What are likely misinterpretations of the results and what can be done to prevent those misinterpretations? |

Table 2. Example of ethical questions to be integrated into teaching Machine Learning ( in Saltz et al., 2019, pp. 32:10).

### 5.1.3 Integrating interdisciplinary approaches to consider context of use in AI design

An essential part of our notion of Reflective AI is that it is not only the end-users that need to be reflective in their use of AI, but also designers and developers themselves need to reflect on how they design AI systems. Beyond ethical aspects, discussed in the previous section, this also includes the question of the overall approach to the design and development of AI systems.

There have been increasingly calls for the designers and developers of AI systems to improve them in a way that considers the needs of the users as well as the context in which they are used.

Most prominently, the approaches of human-centric and socially-aware AI (e.g. Shneiderman, 2021; Leslie, 2019; Chatila et al., 2021; Lukowicz, 2020; Shneiderman, 2020; Abdul et al., 2018; Holton & Boyd, 2021; Lindgren & Holmström, 2020; Wang et al., 2020) highlight the need to put people as users and stakeholders (their needs, values and possible consequences using AI), a broader social context of the intended use of AI and its implications at the center of attention, rather than the available data or technological capabilities of AI.

The human-centric aspect is intended as a counterpole to often criticized technology-driven approaches. In its most encompassing form this includes the consideration of ethical, social/societal, legal and environmental concerns and implications for the design and intended use of a given AI system (e.g. Dignum, 2019).

However, the developers of AI can also take into account research from other disciplines, such as psychology or social sciences in order to understand and approach better the context in which users will be using AI systems. The following two case studies contain specific application scenarios that illustrate how integrating interdisciplinary approaches could help 1) fight misinformation by considering the context in which information sharing occurs on social network sites and 2) improving AI algorithms so that they provide more meaningful recommendations for users to achieve behavioral change.





## Case study 1: Addressing the problem of misinformation by considering the context in which communication occurs on social network sites

The problem of **misinformation on social media** has been approached as a problem of content moderation. The traditional role of the editors of a newspaper which decides what gets published or not is now replaced by algorithms that scan user's posts on social media, compare them against a database of known hoaxes and flag them. This solution is not enough to deal with the deluge of misinformation out there because it treats information as an undifferentiated epistemic good and the users as epistemic agents. Unless we refine the existing algorithmic approaches to misinformation on Social Networking Sites (SNSs), we risk censoring people and missing out on the disinformation with genuine harmful effects.


In solving the problem of misinformation on social media we need to understand the particular weak epistemic context in which users are acting (Marin, 2020). Users do not post or share (mis)information primarily to inform others, rather many try to make up their own minds of what they should believe by testing how their followers respond to their posts. We are social creatures who decide what to believe based on our social ties with others: if the majority goes one way, very few of us will choose the opposite way. SNSs allow for a quick sample of what others think by allowing users to post an item of news (be it information or misinformation) and then gauging how others react and then making up their minds. In this circumstance, posting and sharing have an epistemic function but only after the post has been reacted to.

Thus, if we look at posting and sharing as speech acts, users do not necessarily assert what they share (Rini, 2017) i.e. they do not claim that it is true - rather they make a gesture of pointing at something (Marsili, 2020) seemingly saying "look at this, I find this interesting, what do you think?" Thus, the social media traffic and user-generated content is similar to a large conversation in which people point at things and then decide later if they believe or not. This conversational pragmatic aspect cannot be addressed by current algorithms that aim to detect false content from truthful ones. Yet the conversational context is what decides the difference between a toxic piece of disinformation and a mildly misinforming news-piece meant to stir conversation.

Existing algorithms cannot pick up the conversational context and the user's intentions yet. The context of the utterances on SNSs has several very specific features that need to be taken into account. Primarily, it is weakly epistemic (Marin, 2020): meaning that users are not necessarily aiming to inform others or be informed, yet the informative effect happens in the background when users get to know about things they did not intend to.

Users act as inadvertent informers to their followers, even if perhaps their intention when posting was of irony, sarcasm, or stirring a debate. Secondly, it is highly emotional: social media uses emotional expressions as shortcuts for meaning (think of the emoji as reactions, the likes and the hearts, that replace spoken language) and users come to seek emotional validation on SNSs.

Therefore, we need to understand the misinformation shared and posted on SNSs as moves in a conversation charged with emotions where people mirror and respond to other's emotions more than to their own content (Marin & Roeser, 2020). These two contexts are only some of the most obvious ones, but there are multiple other ways in which context on social media is different from the mass-media context or that of face to face communications. Hence, future research for Reflective AI should ask **how are the conversational contexts specific to social media, how many distinct contexts are there, and how could these be detected by AI?**



To begin tackling the problem of the conversational context on SNSs, one would need first to outline the types of conversational contexts on social media (such as emotional, epistemic, normative, playful, performative, experimental, etc.) and then devise methods for detecting those. AI algorithms would need to be trained on large sets of user posts to detect this context and classify it. After this step, research needs to look into possible ways to nudge users or make them aware of the context that they are using and how opaque this may be to other users. What we imagine to be clearly ironic or sarcastic may not be perceived thus by the readers of our posts and miscommunication occurs frequently when we only read other's words without seeing their body language or hearing their tone of voice. Reflective AI could also look into how to supplant the lack of embodiment in communication by putting in place markers and symbols that make the conversational context clear to other users.

### Case study 2: Accounting for user-specific factors when providing behavioral change recommendations

When the recommender systems are used to help users to change their behavior when they are not satisfied with their current behavior, traditional approaches might be less effective. As the user is not satisfied with the current situation, building recommendations on historical data is suboptimal (Ekstrand & Willemsen, 2016). We therefore argue that there is a need for novel recommender methods that take this into account. One solution could be to filter recommendations based on specific user goals. For example, food recommender systems built on existing data sets often recommend unhealthy recipes as those are typically the more popular ones on the platforms (Trattner & Elsweiler, 2017). Trattner and Elsweiler show that postfiltering the recommendations based on nutritional scores (like the FSA score used in the UK) can improve the healthiness of the recommendations. Similarly, other approaches that use digital nudging (Jesse & Jannach, 2021), esp. when personalized to the user, might be successful in helping users to improve their behavior.

However, these approaches do not have an underlying model of behavioral change and do not take into account that what to change might strongly depend on the users' ability to do so. One approach that can do this is based on the Rasch scale, which was originally used to measure (environmental) attitudes based on actual behavior of people, rather than their stated attitudes or behavioral intentions (Kaiser et al., 2010). The Rasch scale orders items based on their behavioral difficulty, and matches these with the ability of the user to provide recommendations which are relevant but still achievable. This method was shown to be effective in energy recommendations (Starke et al., 2017, 2020), blood pressure management (Radha et al., 2016) and Food Recommendations (Schäfer & Willemsen, 2019).

The basic premise of a Rasch recommender is that users are provided with measures that are challenging but still attainable, rather than items that are too general and too easy or on the other hand very difficult. For example, in the food recommender, rather than recommending to improve the worst performing nutrients (which are often the difficult ones to achieve) the system recommended to improve the ones that were most likely the ones users could still change. Moreover, the Rasch scale often ranks very different behaviors on the same scale: in the blood pressure management study, we find that measures such as exercising were mixed with measures to reduce salt intake or diet changes. Easy and more difficult measures of each type can be found across the scale allowing to recommend diverse and effective measures to all patients.

This approach can be also taken when e.g. users want to change their technology addiction or any other patterns, and thus the recommendations can be employed to stimulate productive behaviors. Technology addiction is a serious problem that has emerged not so long ago (D'Arcy

at el., 2014) especially on social networks (Serenko & Turel, 2015). In-depth understanding of the properties that triggers technology addiction would help to design Reflective AI systems already from the start. We propose an approach for designing Reflective AI systems which takes the hidden learning outcomes of systems into consideration. Analyzing and understanding the essentials of what systems really teach people, how they really affect people is the first step towards designing Reflective AI systems.

To sum it up, to deliver fair and explainable recommendations an integrated solution is needed: the development of the right recommender algorithms is just one piece of the puzzle, and is part of a larger (eco) system of supporting actors. Rutjes et al. (2019) have argued that lifestyle coaches often hesitate to use data and apps in their coaching practice, showing that there are several barriers to actually implement these type of systems into the daily coaching practice, stressing the need for a value sensitive design and user participatory design approach (Ekstrand & Willemsen, 2016).

## 5.2 Organisational practices for Reflective AI

The previous section outlined changes needed in the existing working practices of AI developers and designers. Here, we go one level further and address overall changes needed in organisational logics and structures in order to foster the development and implementation of Reflective AI technologies and practices. In this chapter we address two main components:

- 1) Integrating Reflective AI in organisational innovation adoption,
- 2) Changing values of commercial organisations.

The following diagram illustrates possible solution approaches in terms of organisational practices, structures and processes (circled in green) that will be addressed in the next sections.

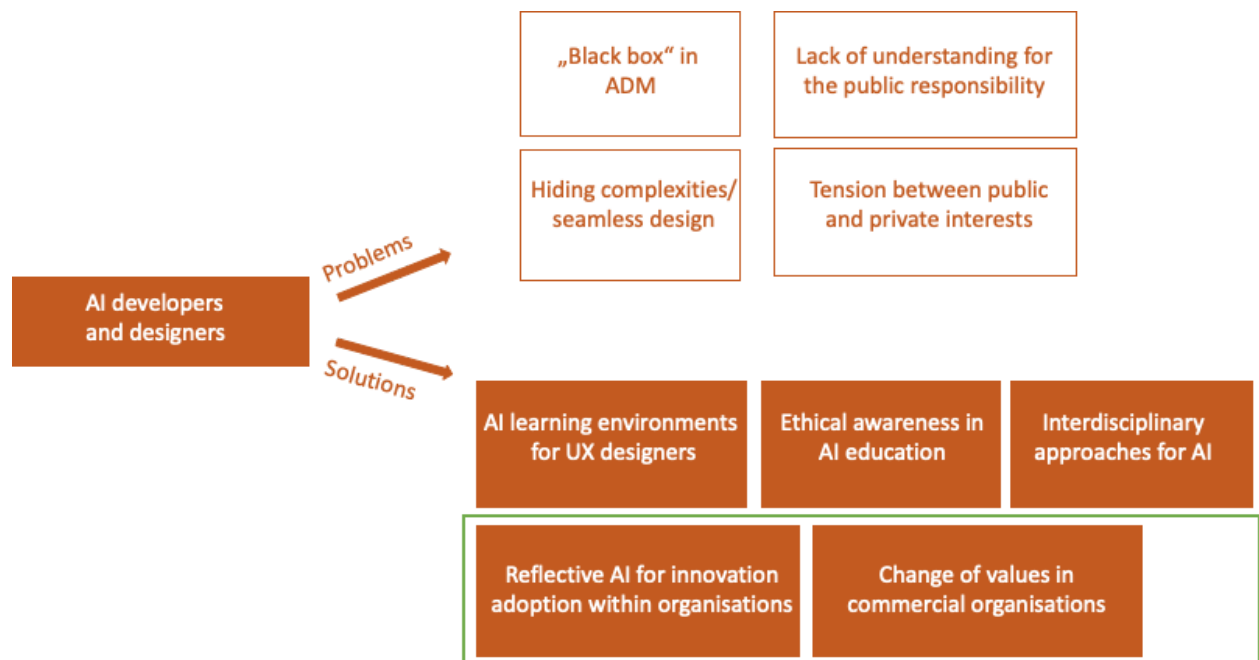



Diagram 3. Towards Reflective AI: Problems and solution approaches regarding AI developers and designers in terms of organisational practices

### 5.2.1 Integrating reflective AI in organisational innovation adoption

The rapid digitalisation in recent years poses a challenge for all types of organisations - governmental, non-governmental, administrative or corporate - to adapt their operations and



internal processes according to the emerging digital trends, especially in the AI sphere. Examples include the integration of electronic filing systems, the emergence of AI-prepared company reports<sup>10</sup>, legal and other texts, as well as the adoption of automated decisions (e.g. for marketing goals). All such organisational innovation adoption processes require intellectual, strategic and political reflection, review, interpretations and organizational contextualization as well as possible adjustments. As such, organizations need to explore AI systems by addressing first and foremost the interdependencies and interactions between employees and managers within the given organizational structure in the context of digital innovations and in particular in view of the increasing application of AI.

At the same time, the users and decision-makers within the organizational structure and hierarchy need to retain their sovereignty of interpretation and development of AI to arrive at Reflective AI systems. However, this is often difficult to achieve because within organizations and their internal cultures, the effects of AI systems on human decisions and actions are still insufficiently recognizable and often incomprehensible for most of the concerned actors. Consequently, there are not many ways in which organisational employees can offer or formulate their digital needs for AI services.

Solutions in this regard need to be based on the adoption of a holistic and differentiated exchange between AI developers, AI users in the broad sense and their identified needs in the respective organizational setting which includes the consideration of existing IT technologies already in use. One way of achieving this could be to embed a human-centered development and learning laboratory on reflective artificial intelligence, in short RAI-LAB, within the organisational structure. This lab should be an integral part of a respective organization and act as a learning and developing entity for the entire organization, its employees, its programmes and processes, decision-making and strategy development as well as the overall functioning of the organization.

The establishment of a lab like this would require that all employees of a given organisation (teams, leaders, their interactions, patterns/structures) should, therefore, be an integral part of the RAI-LAB in order to participate in the digital and social transformation process of the organisation. In the RAI-LAB approach, research, development and implementation/integration of AI systems takes place in an organisation to test AI systems for their accountability and trustworthiness as well as their impact. The organisational impact assessment of deployed AI systems is jointly reflected, reviewed and adjusted from different perspectives (difference-oriented). The transformation of social conditions (communication, decisions, contexts) is given high consideration.

### **5.2.2 Value changes of commercial organisations**

In order to provide for transparency, fairer recommendations or to ensure user privacy, companies which employ AI algorithms to provide services to their customers often report that they experience trade-offs with their existing metrics, such as lower levels of engagement or reduced convenience for the users. For example, some media company representatives we interviewed use algorithms that rerank and boost content which has higher public value, in order to provide for the diversity of the recommendation set. As a result, their recommendations become less homogenous and the engagement of the users decreases. In a similar vein, in order to provide a targeted recommendation, companies often collect demographic data, to be able to better match the users and to identify their needs, or in order to make an easy and convenient log in, they offer authentication through Facebook, thus automatically sharing the user data with a third-party service (for a broader overview on the privacy issue see section 3.1.2 and 4.2.3).

Thus, on one hand, in order to be fair, transparent, and provide for explainability, a company needs to consciously adopt these trade-offs in its company policy and support and stand behind

---

<sup>10</sup> See PR 20/20: <https://blog.hubspot.com/marketing/how-to-shrink-reporting-time-with-ai>



them. Although the engagement rates might get lower or the recommendations might be less exact, they ensure the fairness and transparency of the system provided to the end-users. This, in turn, can have a good impact on the relationship with customers, if the latter see that the company has values different from pure profit maximization. We already see a lot of companies who are adopting this kind of view and, in fact, not compromising the profitability as a result. As one interview partner put it: *“In Europe, they are trying to create a narrative to increase trust and then also to increase profitability”*. Consequently, such values need to be institutionalized in the company and promoted among its employees and also transmitted to the end-users.

On the other hand, it is important to increase awareness of companies of solutions that overcome such trade-offs and demonstrate that **it is a false dilemma that using AI is at odds with values such as transparency and privacy**, e.g. that minimizing personal data requirements needn't compromise the value for the users (see also Chapter 3.4). Moreover, as customers attach more importance to such human-centric values, companies need to reconsider the evaluation metrics they use to measure customer engagement and satisfaction. The development of evaluation metrics which consider not only the accuracy or click-through rate, but also more human values such as critical thinking, trust, bias and fairness is crucial. Existing research on developing such metrics shows both the challenges and the way forward (Chouldechova & Roth, 2018).

**5.3. Structural changes for Reflective AI**

As already stated, Reflective AI is a holistic and comprehensive approach that acknowledges the need not only for individual and organisational changes, but for broader societal and structural shifts in order to create and use AI technologies in a way that harnesses their benefits. The role of governments, international organisations and supra-governmental structures (e.g. the EU) to control and audit the creation and deployment of AI technologies, as well as to ensure that citizens have access to proper educational possibilities to learn about AI is crucial. The structural changes needed to establish the notion of responsible and reflective AI development and use are complex and need to address different areas, however, in this report we are focusing on two main aspects - auditing and literacy - as they were outlined as the most pressing issues by many of our interview partners.

The following diagram summarizes the main problems that public institutions (e.g. AI regulators) face when dealing with AI technologies (as outlined in Section 3.3). It also illustrates possible solutions in terms of institutional and structural changes that will be addressed next.

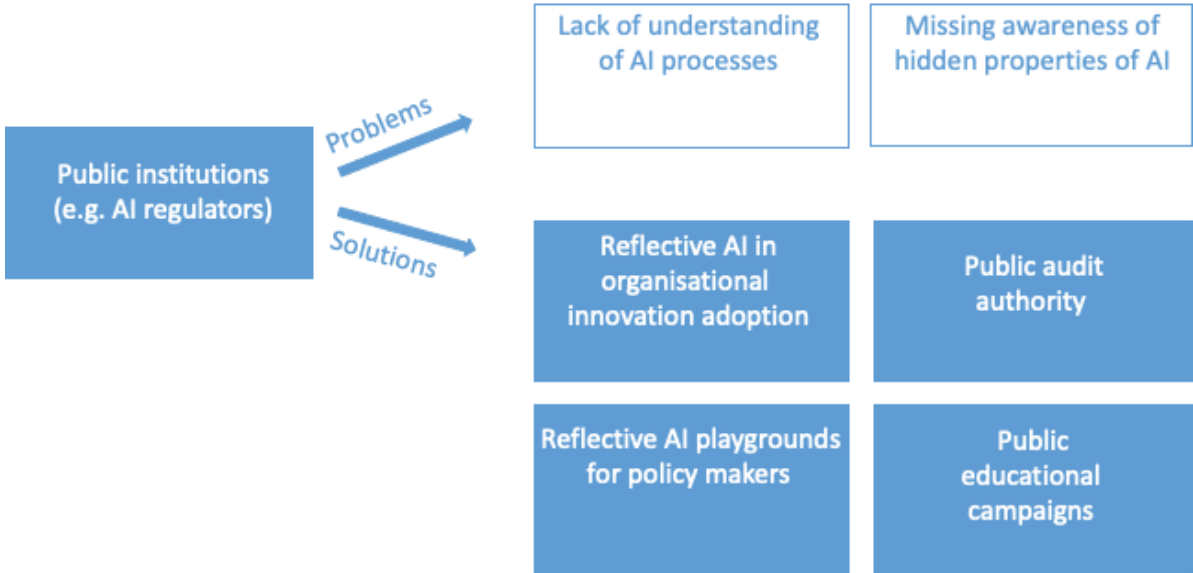


Diagram 4. Towards Reflective AI: Problems and solution approaches regarding public institutions



### 5.3.1 Auditing and control of algorithm development and deployment

As already outlined in section 3.3, one of the key issues according to many interviewed experts is the slow and insufficient governmental control over AI development and deployment. Even if there are some significant steps towards achieving a comprehensive regulation over private data use by companies (GDPR), there are still many further aspects that need to be better regulated.

One idea expressed within the expert interviews was the establishment of an audit authority which would define compliance criteria for AI systems and would check whether the services and products that employ AI comply with them. These compliance criteria would be non-negotiable, especially for the high risk and high impact applications. In this way, the burden of evaluating and being informed about possible consequences of AI which is currently with the end-user would be relieved and the developers would be additionally incentivized to develop systems which are less discriminating and less biased.

The implementation of such an authority and the definition of the compliance criteria as well as the methods for checking them are far from trivial, because the “one size fits all” approach would hardly work for all types of actors involved in the development of AI algorithms. An additional question would be by whom such a controlling entity should be operated (governments, civil sector) and how legitimate will it be. Currently, the EU commission is already thinking of ways to organize such an authority and respond to such calls for more control. An important step of the Commission in this direction is the proposal on banning the use of AI for mass surveillance and/or ranking behavior (like the “social scoring” in China) (Chee, 2021).

### 5.3.2 AI literacy and public education about AI

In 5.1.2 we tackled the need for a better educational curriculum for AI developers and designers. However, there is also a necessity to educate the general population about basic principles and properties of AI (see 4.2.2) or about the risks that unreflective AI use poses (as outlined in 2.1). In order to reach as many people as possible, educating citizens about AI should be a large-scale collective and well coordinated effort.

Therefore, in order to shape public opinion, governments could issue mass AI educational campaigns to demystify such technologies and explain how they work. Such educational campaigns, programmes and clips on new technological appliances were done, for example, in the 80s by the BBC<sup>11</sup>. Nowadays they could be done, for instance, through trusted social media channels, or through government-sponsored MOOCs. One example could be projects such as *AI Competence for Sweden*<sup>12</sup> - a national initiative for education and competence development in artificial intelligence for working professionals.

However, it is important that such campaigns are created in a way that reaches all segments of society and not only people with higher education and from a privileged socio-economic background. Educational projects like *Elements of AI*<sup>13</sup> have the vision to bring AI closer to the general public and make these systems more understandable to everyone. *Elements of AI* is not active only within one country, but the contents from the online courses have been translated into many different European languages thus ensuring that people across the European space are better educated about AI technologies. Such initiatives coordinated on national and global level should be further supported by both national governments and other (public) institutions.

Furthermore, interview partners were advocating for more AI literacy opportunities already in the curriculum in primary school or high school. By this they did not necessarily mean to teach children new technical competences (e.g. how to code), but to teach them to be able to


---

<sup>11</sup> <https://www.bbc.co.uk/taster/pilots/computer-literacy-project>

<sup>12</sup> <https://ai-competence.se/en/>

<sup>13</sup> <https://www.elementsofai.com/>





understand how digital technologies work and ask critical questions about such phenomena. Additionally, some experts were suggesting integrating AI literacy courses also in university education. As AI is becoming all-encompassing, being integrated into many daily activities, it needs to be understood not only by the future developers of AI, but also by other specialists, such as UX designers, product managers etc. and courses on AI should be included in the curricula of many other discipline majors as part of general education on the subject matter.

## 6. Directions for further research

*This section synthesizes the main challenges and directions for future research related to the vision of Reflective AI based on insights from Sections 2-4. What needs better understanding? What are the blindspots? What should new approaches consider? What streams of research should be connected?*

The challenges and research directions identified in the previous sections fall into two areas:

- How to design systems and solutions enabling a reflective use of AI?
- How to create enabling work practices and organisational conditions for Reflective AI development and design?

For an overview, the main problems and research directions in each of these areas are first summarized in Table 3 and Table 4 below. The subsequent sections describe them in more detail. This synthesis follows the same leading questions that have guided this report on what needs to be better researched for ensuring a reflective use and development of AI.

| CHALLENGE: Designing systems and solutions for reflective use of AI  |   |
|--|---|
| Problems/user needs  | Directions and questions for future research  |
| <p><b>Transparency of AI presence</b><br/>Lack of transparency: end-users don't know that AI technologies are in use</p>   | <ul style="list-style-type: none"> <li>• How to signal the presence of AI technologies in an engaging and understandable way, so that users' attention is attracted towards the fact that AI technologies are in use, but without overloading the users with too much information?</li> </ul>   |
| <p><b>Understandability of AI</b><br/>Lack of understandability for the key operational principles of AI technologies</p> <p>No "one-size fits all" explanations: not all provided explanations for the inner principles and properties of AI are suitable for people from different user groups</p>                         | <ul style="list-style-type: none"> <li>• What are the most important properties of AI that should be understood by users to allow competent and reflective use of AI?</li> <li>• How could hidden properties of AI be exposed and made understandable to the users?</li> <li>• How could this be achieved so that users internalize this understanding in new, more appropriate mental models of AI, its benefits and risks it carries?</li> <li>• How to devise explanations of operational principles and properties of AI that are comprehensible for a wide-range of users, while sufficiently precise to set the ground for understanding subsequent explanations of potential risks?</li> </ul> |
| <p><b>Diversity and "birds-eye view"</b><br/>Lack of a "birds-eye view": users see only the personalized results presented to the by AI recommendations, but not the whole picture</p> <p>Many of the current techniques in recommender systems don't sufficiently account for diversity in the recommendations provided</p> | <ul style="list-style-type: none"> <li>• How could AI systems (e.g. recommender systems) inform the users where they stand with regards to other users?</li> <li>• How could personalization be balanced with an awareness of a diversity of possible views, without overwhelming the users?</li> <li>• How can diversity and perspectives in recommender systems be defined and measured (e.g. in news recommendations or in the selection of posts in social networks)?</li> <li>• What normative considerations are required to ensure transparency between personalization and a birds-eye</li> </ul>   |



|   |  |
|---|--|
|   | <p>view for users?</p> <ul style="list-style-type: none"> <li>• How could such principles be translated into design decisions that satisfy user needs (e.g. relevant content)?</li> <li>• How should AI systems give users effective autonomy and control over the level of personalization they desire?</li> </ul>  |
| <p><b>Control over use of personal data by AI</b><br/>End-users are often not aware about actions they can take online in order to secure their data privacy when using AI technologies</p> <p>Many of the existing approaches in developing and designing AI compromise user privacy</p> | <ul style="list-style-type: none"> <li>• How could the underlying principles of privacy-preserving techniques and their implications in practice be explained to a wide-range of users and stakeholders?</li> <li>• How can we design solutions that protect individuals, but still allow companies, governments and society to harness the benefits of big data and AI?</li> </ul>  |
| <p><b>Experiential learning and reflective AI experiences</b><br/>End-users lack opportunities to experience the effects of AI technologies in ways that allow experiential learning about the properties and principles of AI</p>  | <ul style="list-style-type: none"> <li>• How could new user experience design patterns for AI systems enable more reflective use of AI?</li> <li>• How can interactive environments for experiential learning about AI be designed and implemented?</li> <li>• How can situations be created which allow end-users to experience the behaviour of AI systems and their possible individual and societal consequences?</li> </ul> |

Table 3. Challenges and research directions for systems and solutions for reflective use of AI

| CHALLENGE: Creating work practices and organisational conditions for Reflective AI  |   |
|---|---|
| Problems/needs  | Directions and questions for future research  |
| <p><b>Work practices in AI design &amp; development</b><br/>User-experience designers lack knowledge about the inner workings of AI technologies</p> <p>AI developers often lack awareness of ethical issues and potential harmful effects connected to the technologies they develop</p> | <ul style="list-style-type: none"> <li>• How to develop AI learning environments and possibilities for user experience designers?</li> <li>• What concrete designers' needs should be addressed thereby?</li> <li>• What are the best strategies/ways to sensitize AI developers, machine learning students etc. about the ethical implications and responsibilities of their work?</li> </ul>  |
| <p><b>Adoption of AI in organisations</b><br/>Organisations that integrate AI technologies in their internal operations need mechanism to do so in a way that allows employees to be an integral part of the innovation adoption process</p>  | <ul style="list-style-type: none"> <li>• In which way organizations need to develop in terms of structure and human competencies when their overall functioning and decision-making processes are increasingly dependent on AI systems?</li> <li>• What adaptation is required from organizations with regard to their social interacting systems, core functions and the embedded organizational contexts?</li> <li>• How can effective control be ensured in an organisational context so that AI systems act in a responsible, transparent and responsive manner?</li> </ul> |

Table 4. Challenges and research directions regarding work practices and organisational conditions for reflective AI



## 6.1 Demystifying AI: Transparency, Understandability, Diversity, Control

To develop effective approaches for demystifying AI, existing misconceptions of AI held by different types of actors need to be better understood (e.g. users in private contexts, decision-makers in professional use, policy-makers). General public perceptions of AI and misconceptions of specific types of AI systems are increasingly being studied, especially from the perspective of human-computer interaction (e.g. Eslami et al., 2016; Alizadeh et al., 2021). Particularly relevant are studies of users' mental models of AI and how these are related to system affordances (e.g. Devito et al., 2018; Eslami et al., 2016; Hernandez-Bocanegra & Ziegler, 2021). But how to support the development of suitable mental models of AI has so-far been little addressed (Kulesza et al., 2013).

Investigating mental models users have of different types of AI systems should **identify design considerations and system affordances that need to be addressed to allow people to form correct mental models of AI. Achieving this will enable both a safer and a more productive use of AI and its benefits.** This research should be undertaken in interdisciplinary teams that can both uncover the underlying psychological and social issues in the formation of mental models in human-AI interaction, and propose concrete design solutions and guidelines to address them.

Some general principles from existing knowledge in human-computer interaction will likely apply to human-AI interaction, but specific considerations will be needed for different types of AI in different contexts of use. In particular, this concerns the role of social context and social interactions in the formation of mental models and theories about AI (e.g. “folk theories” (Devito et al., 2018)) where few substantial findings are available so far.


We believe that the four levels of affordances that we have highlighted in this study (Chapter 4) can provide some general orientation, but how exactly they can be best put in practice is still a widely open question that requires much further research. Some of the main challenges and research directions in this regard we summarize below.

### *Transparency of AI presence (“AI inside”)*

The need for transparent signalling of the use of AI in a given system to its users has already been highlighted in some research (Hamilton et al., 2014) and normative guidelines (see Fjeld et al., 2020). But what level of detail this signalling should provide (e.g. just in general vs. specific functionalities) and with what type of information (e.g. purpose, effects) are still open questions.

In Chapter 4.1 we have proposed several different levels of signalling for ensuring that users can form a meaningful awareness about the role, purposes and effects of the use of AI in a system. But how these different levels of signalling of AI presence should be provided, so that they attract user attention and avoid information overload, are easily understandable and engaging are all open and challenging questions for further research.

Some of these challenges are related to psychological factors determining user acceptance of explanations of AI results (see review in Wang et al., 2019). Other relate to experiences from previous work on designing interactive systems that stimulate reflection and behavioural change (e.g. in health (Kocielnik et al., 2018b), learning (Kocielnik et al., 2018a) or pro-environmental behaviour (Novak et al., 2018; Koroleva et al., 2019; Böckle et al., 2018)), and consider the role of social interaction in doing so (e.g. Ploderer et al., 2014). The form in which such explanations should be provided is closely related to research on different types of explanations and their presentations from human-centric approaches to explainable AI (e.g. Wang et al., 2019).



Research in explainable AI has also already shown that different types of users may require different types of explanations for different purposes (e.g. Bhatt et al., 2020). But since AI is often used in wide-scope systems serving very different types of users (e.g. search engines, social networks, recommendation systems), explanations of the presence, purposes and effects of AI in such systems cannot be provided in the same way, at the same level of detail for all users.

This points to further research on (user-controlled) adaptability of explanations of AI presence. This could include techniques such as scaffolding (e.g. from computer-supported learning), that allow different levels of complexity to co-exist and be uncovered progressively without overburdening the user (Jackson et al., 1998; Sharma & Hannafin, 2007).

Further research in this area could thus benefit from building on existing work in algorithmic awareness (e.g. Alvarado & Waern, 2018; Eslami et al., 2015; Lee et al., 2019; Hamilton et al., 2014), explainable AI (e.g. Wang et al., 2019), human-AI interaction (e.g. Amershi et al., 2019; Zang et al., 2020) and persuasive communication for behavioural change (e.g. De Wit et al., 2008; Moyer-Gusé, 2008; Novak et al., 2018; Koroleva et al., 2019).

Finally, as the transparent provision of different levels of information about the presence and purposes of AI use in a system depends on the willingness of companies to provide it (which in turn depends on their business models), this research should also consider regulatory aspects (e.g. mandating disclosure through law) or other forms of incentives (e.g. providing transparency of AI presence to increase user trust).


#### ***Understandability of operational principles, properties and risks of AI***

Establishing user awareness of AI presence and the purposes of its use in a given system is only a starting point, not the final purpose. To fully empower a reflective use of AI by end-users requires them to develop a better understanding of what AI is, how it operates and what effects and risks its use can result in. **An overarching research question we see here is: What is the level of explainability that is required by end-users to understand the main workings and consequences of AI systems, so that these can be used reflectively?**

AI models that are interpretable by design are a prerequisite for reliable explanations that different types of users and stakeholders can understand. Post-hoc explanations of black box machine learning models are often unreliable and can be misleading even for AI experts (Rudin, 2019; Rudin & Radin, 2019).

Combining research on interpretable machine learning (e.g. representational learning) with research on human-AI interaction carries the promise of developing new solutions for trustworthy AI systems that are verifiable by experts and whose workings and consequences can be appropriately explained to lay end-users and stakeholders. Ensuring interpretability is also required for showing how the internal workings of AI models relate to both expected benefits and potential risks. Uncovering and making such relationships observable is crucial for enabling critical reflection.

We have proposed that one way to address this is to make the key hidden properties and risks of AI understandable to end-users. A large body of work has already investigated how different types of explanations of results of AI systems can help users develop some understanding of why a specific AI system has produced a specific result in the given situation (see e.g. (Miller, 2017; Abdul, 2019; Wang et al., 2019) for an overview). But research on how end-users can be enabled to understand the underlying properties of AI (e.g. sensitivity, temporal effects) and their consequences, is to the best of our knowledge in its infancy.



Accordingly, open questions for further research abound. This starts with diametrically opposing views of whether such an understanding can be acquired by end-users without proper formal education. As argued in Chapter 3, we acknowledge that expert-level understanding of AI systems cannot be expected from “laypeople”, since even for AI developers the complexities involved can be daunting.

We propose that further research could and should aim at identifying key properties of AI systems that, if exposed to users in appropriate ways, can help them grasp both the underlying nature of AI, its benefits and possible risks involved in its unreflected use. **We have proposed five such key properties of AI: sensitivity of AI algorithms, non-linearity and temporal effects, the “birds-eye view” and privacy preservation.** But there are bound to be others, possibly depending on specific classes of AI techniques or contexts of use.

Some questions for further research thus include: **What are the most important properties of AI that should be understood by users to allow competent and reflective use of AI? How could hidden properties of AI be exposed and made understandable to the users? How could this be achieved so that users internalize this understanding in new, more appropriate mental models of AI, its benefits and risks it carries?**

These are highly interdisciplinary challenges. Research in various fields has shown that the effectiveness of information or explanations about complex issues or phenomena depends on many factors, such as the compatibility with existing beliefs and opinions (Knobloch-Westerwick et al., 2020), the message style or narrative framing (e.g. De Wit et al., 2008).

This illustrates another major challenge: **How to devise explanations of operational principles and properties of AI that are comprehensible for a wide-range of users, while sufficiently precise to set the ground for understanding subsequent explanations of potential risks?**

On one hand, promising avenues for further work could include integrating interpretable machine learning with research on narrative strategies from persuasive communication (e.g. Slater, & Rouner, 2002) and with existing work on human-centric perspectives on explainable AI (e.g. Miller, 2017; Wang et al., 2019). Lessons from behavioural change and communication regarding health risks or pro-environmental behaviour also suggest that using negative messaging to highlight risks is less effective than positive messaging. Accordingly, **solutions for exposing hidden properties of AI and their relation to potential risk should also address the expected benefits of AI in a given system.** If explanations are used as a method of addressing this challenge, solutions need to be found that make such explanations relatable to the user, to their current experience and current context.

In this area, a promising avenue for future work are **interactive explanations that allow users to actively construct their understanding of the system operation and its underlying properties**, along the lines of constructivist theories of learning (Ackermann, 1996). This could expand existing work on interactive recommender systems (He et al., 2016; Jugovac & Jannach, 2017) and interactive machine learning (Dudley & Kristensson, 2018), that has already shown how interactivity can provide important benefits in users’ understanding of AI.

By interactively engaging with the system, users would not only understand it better, but also be better able to consciously decide if they are willing to use the system at all. As learning from experience happens through reflecting on what one has experienced, the design of such solutions could also be informed by the theory of experiential learning and its applications (e.g. Kolb, 1984; Morris, 2019).





## Diversity and “birds-eye view”

Developing an awareness and understanding of possible individual and societal effects of AI use requires the ability to take on a birds-eye view, that shows possible views of the system and its results as it would be experienced by many different users (Chapter 3.1.2). Such views are not available to normal users as the system behaviour and results they experience are often dependent on their preference profiles and previous interaction with the system (Hamilton et al., 2014). That makes it difficult to understand how a system using AI may lead to harmful effects, such as facilitating misinformation or online radicalization (Ribeiro et al., 2020).

**Further research should thus investigate possibilities for allowing users to experience such a birds-eye view, to enable them to grasp how different users may experience very different views of the system and the information it presents them.**

Incorporating such functionalities in the design of AI systems is one way to support an awareness of specific hidden properties of AI and their effects. This leads to research questions such as: **How could AI systems (e.g. recommender systems) inform the users where they stand with regards to other users? How could personalization be balanced with an awareness of a diversity of possible views, without overwhelming the users?**

A case in point is the design of recommender systems for news recommendations with respect to personalization and diversity issues. As AI-driven recommender systems for news recommendation optimize for user engagement and employ collaborative filtering, their recommendations are closely tailored to inferred user interests (Chapter 4.2.2) (Bernstein et al., 2020). This reduces both the diversity of information and the awareness of available perspectives. The bird’s eye view is missing.

This relates a number of existing research challenges to the goals of Reflective AI. On one hand this research can build on existing work on interactive and diversity-optimizing recommender systems (e.g. in the news domain Vrijenhoek et al., 2020). This includes challenges such as: How can diversity in news recommendation systems be quantified in accordance with normative considerations? How should diverse content be integrated in recommender settings?

This is additionally complicated by both psychological factors and existing user expectations that have been formed through their experience of existing highly personalized systems (e.g. perceiving diversity in recommendations as poor performance or paternalistic (Bernstein et al., 2020)). However, addressing these issues is not just a technical challenge. Normative considerations regarding diversity in sources and perspectives are also difficult to define and still missing.

Thus, difficult challenges in providing a “birds-eye view” to facilitate a more reflective use of AI call for further research. Some of these include: **How can diversity and perspectives in recommender systems be defined and measured** (e.g. in news recommendations or in the selection of posts in social networks)? **What normative considerations are required to ensure transparency** between personalization and a birds-eye view for users? How could such principles be translated into design decisions that satisfy user needs (e.g. relevant content)?

**How should AI systems give users effective autonomy and control over the level of personalization they desire? And what would people need to understand** about the hidden properties of personalized systems, their individual and societal consequences, **to competently make such decisions?**



### **Control over the use of personal data in AI (“privacy preserving AI”)**

The need to provide human control over AI processes for high-risk applications such as when AI algorithms are used to support decision making with potentially significant consequences (e.g. health, justice, recruiting) has been highlighted in a number of proposals of normative principles for guiding the use of AI (see (Fjeld et al., 2020) for a review). In research, the idea of “human in the loop” has also been investigated as a way to develop better solutions that combine human and machine intelligence.

We propose that the idea of user control should be expanded as a general principle, especially with respect to the use of personal data that are often used in AI applications. **AI systems should always allow users to effectively control whether and to what extent to contribute or allow access to personal data.** That is both a foundation for user trust and a prerequisite for building an understanding of the underlying workings of the system and the consequences of its use.

On one hand, this requires **research in new approaches for explaining how different types of AI applications use personal data and the consequences thereof.** In particular, the existing implementations of GDPR-compliant information and options for restricting the collection and processing of personal data are problematic because they are difficult to understand and overwhelming for users. **Real user control can only occur if the system has adequately explained its workings to the user, the purposes of using personal data by AI - and the benefits and consequences of this use.**

In particular, transparency regarding possible actions is needed for users should they perceive a system as not being fair or discriminating against them in the treatment of their data. **Providing users with more control over the functioning of AI systems (human-in-the-loop)** could also provide new opportunities for feedback loops between end-users and system developers and support a more human-centric development and improvement of AI systems.


In order to enable users to really understand the consequences of their actions, **future research should investigate how complex bureaucratic and technical texts could be replaced with examples of concrete effects of specific privacy choices on system results and behaviour.**

This would make it much easier for users to understand the stakes involved in a given case and make informed choices. Applying techniques from AI explainability (e.g. counterfactual and contrastive explanations) and combining them with strategies from storytelling and persuasive communication seem promising avenues for that kind of research.

**Most users, companies and policy-makers are unaware that privacy-preserving techniques for AI exist** that can protect personal data while allowing AI applications that require them to safely and securely process them. **Educating companies, researchers, general users, decision makers and policy makers alike, about the possibilities of privacy-preserving AI** and the principles of their operation could dramatically **shift the wrong perception that surrendering privacy is a necessary sacrifice** for taking advantage of AI benefits.

Helping users, AI developers, system providers and regulators understand and apply the principles and possibilities of privacy-preserving AI could help overcome the current binary choice of “*opt-in or don’t use it*” users unwillingly face in many AI applications.

Future research should investigate how the awareness and understanding of the possibilities of privacy-preserving techniques could be best supported, in spite of their technical complexity: **How could the underlying principles of such privacy-preserving techniques and their implications in practice be explained to a wide-range of users and stakeholders?**



Rather than viewing privacy and AI as a dichotomy, more AI research is needed that asks: **How can we design solutions that protect individuals, but still allow companies, governments and society to harness the benefits of big data and AI?** This includes further research on approaches that minimize personal data requirements and allow end-users themselves to protect their privacy by altering data in ways which do not decrease its value for AI applications (e.g. Choi et al., 2017).

## 6.2 Designing for experiential learning and reflective AI experiences

One approach to enabling users to be more reflective in their use of AI, could be to completely **rethink the entire user experience design for AI systems**. Rather than considering AI transparency, understandability and support for reflective use as add-ons, the entire system should be designed from the outset with these goals in mind. For example, **user experience designers could create new design patterns** to visualize and reflect properties such as sensitivity or uncertainty not only when displaying AI results to the user, but in a way that is inherent to every step of users' interaction with the system (e.g. from formulating a query, to receiving recommended results, to analysing and re-adjusting them based on obtained insights).

Reflection is typically triggered by encountering an inconsistent experience, a problem that cannot be solved in the usual way (a breakdown (Baumer, 2015)). But AI systems have become so user friendly (problem-free) that they no longer invite such reflection. Future AI designs should thus consider integrating ideas of so-called “seamful design” (Chalmers & Galani, 2004), where rather than providing a seamless experience by hiding system complexity from the users, the user interface purposefully highlights possible irritations as triggers for reflection (e.g. Chalmers & Galani, 2004; Inman & Ribes, 2019).

For example, such reflection triggers could be provided when system results are uncertain or highly sensitive to small changes in training or input data, or when the consequences of taking them at face value could negatively impact other people. This line of research could also benefit from previous work on interactive systems for supporting reflection (e.g. Baumer et al., 2014; Baumer, 2015; Karyda et al., 2021) and behavioural change (e.g. Novak et al., 2018; Koroleva et al., 2019; Böckle et al., 2018 ).

On the other hand, learning about key properties of AI systems and reflecting on their effects on system results and societal risks requires willingness, time, effort and triggers for conscious reflection (Chapter 4.2.2). It also requires mechanisms that allow for experiential learning, i.e. learning through reflection on one's own experience, rather than being educated by an authority. It is thus difficult to expect users to reflect on their experience and understanding of AI, while they are using an AI system to reach their goal, entertain themselves or perform a task.

Accordingly, an approach to address this would be to **create opportunities for experiential learning outside of the use of specific AI systems**. This could take the form of interactive “playgrounds” that support end-users in gaining a practical understanding of the principles, properties and effects of AI through an experiential learning approach, i.e., learning through reflecting on a concrete experience (Kolb, 1984; Morris, 2019). Future research could investigate how such dedicated **interactive environments for experiential learning about AI could be designed and implemented**. Such environments should allow users to grasp the nature of hidden properties of AI and their implications at the personal and societal level. They should enable them to internalize these insights into better mental models of AI systems. In this report we proposed an example approach to how such environments could be imagined (Chapter 4.2.2)



To develop such environments a number of difficult research challenges need to be addressed. Mental models change when users are faced with real experiences and need to relate and compare them to existing models of previous experience (Johnson-Laird, 1983). We argue that pure information-based approaches using explanations (Miller et al., 2017) and teaching about AI fall short because these approaches do not allow people to learn by reflecting on actual experiences. **But how situations could be created in which end-users could experience the possible behaviour of AI systems and their possible individual and societal consequences is a wide-open question.**

On one hand, **interactive simulations of specific types of AI techniques that make their behavior and properties under different conditions easily observable to end-users** would need to be developed. A number of interactive machine learning toolkits or tools that would allow such simulations in principle are available and some examples allow users to explore specific AI algorithms by interactively manipulating their parameters<sup>14</sup>. But they are either not suitable for users without technical expertise, or they focus on teaching technical skills (e.g. Machine Learning for Kids) - and they don't support experiential learning about hidden structural properties of AI and their personal and societal effects.

**Artistic approaches** have also explored engaging people with reflection on societal problems of some AI technologies (e.g. image classification<sup>15</sup>). Work on nudging users towards more reflective online information consumption for fighting fake news<sup>16</sup> and polarization<sup>17</sup> demonstrates the potential of gamification to engage users. But neither allow users to experience the underlying structural properties of AI systems and how these are connected to personal and societal effects.

Further research should investigate how to design such interactive environments that allow users to experience both the key structural properties of AI (e.g sensitivity, temporal effects) and their relation to possible risks of the use of a specific class of AI techniques. For example, by extrapolating samples of user interactions with the system to a longer period and showing what recommendations the use of the system over specific interaction paths could result in.

Moreover, such simulations would need to place the observed system behaviour in relation to known risks and possible impacts on users in real-world contexts (e.g. openness to extremist views (Ribeiro et al., 2020)). And this would need to be done in ways that allows the users to discover and observe such effects in a trustworthy environment which invites reflection.

Existing approaches to explainable AI cannot achieve this, due to framing it as a technical problem, or at best a problem of individual cognitive reasoning about a specific system or result (Wang et al., 2019). They tend to neglect the role of social context in which AI is used in spite of recent studies highlighting its importance (Eslami et al., 2016; Kou & Gui, 2020). And they do not address the possible aggregated effects of individual results and decisions based on them and their broader societal consequences.

Another critical challenge for successful design of environments for experiential learning about AI is the inherent effort and willingness needed by users to consciously engage into reflection on the results and the behaviour of an AI system while using it. The required cognitive effort is in opposition to users' expectations of a frictionless use of such systems, whose very purpose is to reduce cognitive complexity and information overload (Schmitt et al., 2018; Li, 2017). Moreover, people may ignore the explanations if the results reinforce their existing beliefs


---

<sup>14</sup> See projects such as: Machine Learning for Kids: (<https://machinelearningforkids.co.uk/#!/welcome>), Google AI Experiments (<https://experiments.withgoogle.com/collection/ai>), RapidMiner (<https://rapidminer.com/>)

<sup>15</sup> Excavating AI: <https://www.excavating.ai/>

<sup>16</sup> Bad News: <https://www.getbadnews.com/#intro>

<sup>17</sup> Blue Feed, Red Feed: <https://graphics.wsj.com/blue-feed-red-feed/>



(Knobloch-Westerwick et al., 2020) or defer responsibility to AI because that provides immediate gratification (Ryffel & Wirth, 2020). This is especially likely when the presented results, their explanations and system behaviour are inconsistent with the users underlying intuitive understanding, i.e. their mental model of a given AI system.

All of the above are all difficult challenges that invite further research at the intersection between AI research in general, interpretable machine learning, human-AI interaction and various fields from the social sciences such as ethics, social psychology, learning sciences and communication science. The integration of constructivist approaches to learning (Ackermann, 1996; Resnick et al., 2000) and experiential learning (Kolb, 1984; Morris, 2019) can provide valuable insights for creating engaging learning experiences that help people develop an understanding of how AI works and of its potential personal and societal impact.

### 6.3 Work practices in AI design & development


In section 5.1 of this report we identified three areas that are import for the establishment of new work practices in AI design and development to support the creation of Reflective AI technologies: 1) supporting user experience designers in learning about AI, 2) integrating ethical awareness considerations into AI development and teaching, 3) integrating interdisciplinary approaches to consider context of use in AI design. We are suggested several possible ways to address these issues:

- Creating an experiential learning environment where user experience designers can interactively learn about the core principles and properties of AI (as also suggested by Winter & Jackson, 2020)
- Developing a set of guiding questions for teaching AI awareness in machine learning courses (as also suggested by Saltz et al., 2019)
- Integrating human-centred and interdisciplinary approaches towards AI technologies to address pressing societal and individual issues such as the spread of misinformation online or the development of comprehensive recommendations based on the user's needs.

These initial ideas and suggestions call for extended further research. For instance, future research is needed to understand how exactly to develop **an experiential learning environment about AI specifically for user experience designers** and what specific needs of UX designers should be addressed when doing so. Furthermore, the motivations of designers to use such environments and learn more about AI should be researched in more detail to understand better how to keep them engaged in such environments and provide for the best learning outcomes possible. If such experiential learning settings exist, their effectiveness as well as the effectiveness of alternative approaches towards learning should be tested and compared.

Ethical considerations should be integrated as an essential part of AI development and technical AI education. As mentioned in the report, some of the main guiding principles for a responsible design and use of AI have been described in a rising number of documents by different types of actors (for a review see Fjeld et al., 2020).

They include privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human-control of technology, professional responsibility, promotion of human values. However, it should be further researched how these principles could be best and most effectively integrated within the work of AI designers and developers. One important aspect in this regard is the ethical awareness building in AI education. Thus, it should be conceptually and empirically tested which approaches towards sensibilizing students from disciplines such as machine learning are the most effective ones.



Finally, as demonstrated in 5.1.3, interdisciplinary work and approaches are crucial in developing AI technologies that are human-centric and account for the context of use of such technologies. Thus, closer collaboration between researchers from disciplines such as machine learning, computer science, user experience design, psychology, philosophy, social science, history and law will be needed also in the future to address emerging issues in the development of AI technologies. How to best ensure that AI research and development is done in an interdisciplinary setting in the future is thus a pressing question for this field.

#### 6.4 Organisational adoption of AI

We outlined the need for organisational changes in order to ensure that organizations that are integrating AI technologies in their processes consider the needs of the employees and use participatory mechanisms and formats to guarantee that this is happening. Furthermore, we discussed the importance of value changes within companies and the adaptation of their business models in order to ensure that the technologies they are providing to the end-users don't compromise the principles of Reflective AI design.

To tackle some of these issues, we suggest, similar to ideas outlined in 6.3 and 6.2, the establishment of human-centered development and learning laboratories on Reflective AI embedded within the organisational structure. This would enable employees to learn about AI and its reflective use within the context of the organisation they are part of. How to successfully implement such laboratories, what needs to be considered when doing so and how to motivate employees to participate in such formats are all possible questions for future research. This concrete suggestion points towards one possible solution, but there might be other approaches to consider to ensure that organisations are integrating AI in their processes in a reflective manner.

Therefore, it should be further researched in which way does an organization need to develop in terms of its organizational structure and human competencies when its overall functioning and decision-making functions are increasingly taken over by AI systems? What adaptation is required from the organization in view of its social interacting systems (employees, teams, managers, cooperation, communication systems), its core functions (e.g. programmes, processes, instruments) and the embedded organizational contexts? How can effective oversight and control be ensured by the organizational structure and all actors involved so that AI systems continuously act in a responsible, transparent and responsive manner?





## 7. Summary

In this report we have proposed that there is an **underrepresented perspective** in existing research and practice on ensuring a responsible design and use of AI: **the need to empower end-users to use AI reflectively, conscious of both its benefits and possible harms of uncritical use.** To fully achieve and enable that, all the different actors involved in AI design, application and use need to develop such an understanding and reflective practice. The presented analysis suggests **five main observations that can guide further research and practice of Reflective AI:**

**1) The risks of AI stem not only from problems in AI algorithms, but also from the lack of individual and societal understanding of AI potentials and risks of uncritical use of AI.**

Harnessing *benefits* and *preventing harms* of AI cannot be solved alone through technological fixes and regulation. It depends on a complex interplay between technology, societal governance, individual behaviour, organizational and societal dynamics. Enabling people to understand AI and the consequences of its use and design is a crucial element for ensuring responsible use of AI.

**2) AI needs to be demystified in order to overcome the experience gap and reach AI literacy. The mystification and misconceptions of AI threaten its productive and responsible use.**

The experience gap is the difference between the experience that people have with AI on a day-to-day basis and the experience that they need in order to understand AI at the level necessary to enjoy its benefits and avoid its dangers. This applies both to the use of AI in private contexts and in professional work (e.g. decision-makers). Future research needs to understand misconceptions of AI and the experience gap and find solutions to overcome them.

**3) AI models need to be interpretable by design. Interpretability of AI is a prerequisite for an informed understanding and reflective practice by end-users, developers and designers alike.**

Post-hoc explanations of black box machine learning models are often unreliable and can be misleading even for AI experts. AI models that are interpretable by design are a prerequisite for reliable explanations that different types of users and stakeholders can understand. Research on interpretable machine learning combined with human-AI interaction is crucial for trustworthy AI systems that are verifiable by experts and whose workings and consequences can be appropriately explained to lay end-users and stakeholders.

**4) Designing for Reflective AI experiences requires changes in work practices of AI developers and designers. User experience design should make inherent properties and risks of AI models observable (e.g. sensitivity, diversity, privacy), without overburdening the users.**

In spite of a growing attention to ethical issues in AI development (e.g. de-biasing, fairness and non-discrimination), more awareness of the underlying properties of AI is needed in AI development, research and teaching. This concerns in particular the effects of hidden properties of AI on its results and the risks for individual and societal harms. Educating user experience designers about AI is crucial because their work shapes the perceptions and use of AI.


**5) Reflective adoption of AI innovations in organisations requires changes in organisational values and practices, value chains and processes to align with the needs of different actors.**


Apparent trade-offs between commercial goals, the values of the users and the principles of transparency, fairness and explainability, need to be resolved by reconsidering company values and business models. Identifying and realizing AI potentials in organisations requires participative processes that enable the dialogue between different actors (e.g. employees and managers, AI developers and users) about their needs and values in the organizational context. Establishing organisational laboratories for reflective AI experiences can facilitate human-centered development of and organisational learning about AI and its potential for organisations.



## References

- 60 Minutes/Vanity Fair poll (2016). 60 Minutes/Vanity Fair poll: Artificial Intelligence. CBS News: <https://www.cbsnews.com/news/60-minutes-va-nity-fair-poll-artificial-intelligence/>
- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-18.
- Ackermann, E. (1996). Perspective-Taking and object Construction. In Constructionism in Practice: Designing, Thinking, and Learning in a Digital World (Kafai, Y., and Resnick, M., Eds.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Part 1, Chap. 2, pp. 25-37.
- Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access (6), pp. 52138-52160.
- Adamic, L. & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: divided they blog. LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery, 36-43.
- AI HLEG (2019). Ethics guidelines for trustworthy AI, European Commission Independent High-Level Expert Group on AI. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Akhmetova, R. (2020). How AI Is Being Used in Canada's Immigration Decision-Making. Compas: <https://www.compas.ox.ac.uk/2020/how-ai-is-being-used-in-canadas-immigration-decision-making/>
- Alizadeh, F., Stevens, G. & Esau, M. (2021). "I Don't Know, Is AI Also Used in Airbags?". I-com (20 (1)), pp. 3-17.
- Allcott, H., Gentzkow, M., Yu, C. (2019), Trends in the diffusion of misinformation on social media. Research & Politics 6, 2 (2019).
- Alvarado, O., & Waern, A. (2018). Towards algorithmic experience: Initial efforts for social media contexts. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-12).
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-13.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp. 82-115.
- Aslett, L., Esperança, P. & Chris C. Holmes (2015). Encrypted statistical machine learning: new privacy preserving methods. arXiv preprint arXiv:1508.06845.
- Bakshy, E., Messing, S. & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science (348 (6239)), pp. 1130-1132.
- Baptista, G. & Oliveira, T. (2019). Gamification and serious games: A literature meta-analysis and integrative model. Computers in Human Behavior (92), pp. 306-315.
- Baumer, E. P. S. (2015). Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15), pp. 585-594.

- 
- Baumer, E. P., Khovanskaya, V., Matthews, M., Reynolds, L., Schwanda Sosik, V., & Gay, G. (2014). Reviewing reflection: on the use of reflection in interactive system design. *Proceedings of the 2014 conference on Designing interactive systems*, pp. 93-102.
- Belting, H. (2013). *Faces. Eine Geschichte des Gesichts*. Munich: C. H. Beck.
- Bendel, O. (2018). *The Uncanny Return of Physiognomy*. AAAI Spring Symposia.
- Bernstein, A., de Vreese, C., Helberger, N., Schulz, W., Zweig, K., Baden, C., Beam, M., Hauer, M., Heitz, L., Jürgens, P., Katzenbach, C., Kille, B., Klimkiewicz, B., Loosen, W., Moeller, J., Radanovic, G., Shani, G., Tintarev, N., Tolmeijer, S., van Atteveldt, W., Vrijenhoek, S. & Zueger, T. (2020). Diversity in news recommendations. *arXiv preprint arXiv:2005.09495*.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R. , Moura, J. M. F., Eckersley, P. (2020). Explainable Machine Learning in Deployment. *FAT\* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657.
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI 2017 Workshop on Explainable AI (XAI)*.
- Böckle, M., Micheel, I., Bick, M. & Novak, J. (2018). A Design Framework for Adaptive Gamification Applications. *Proceedings of the 51st Hawaii International Conference on System Sciences 2018 (HICSS)*.
- Böckle, M., Novak, J. & Bick, M. (2017). Towards Adaptive Gamification: A Synthesis of Current Developments. *Proceedings of the 25th European Conference on Information Systems (ECIS)*.
- Boden, M. A. (1978). Social implications of intelligent machines. In *Proceedings of the 1978 annual conference - Volume 2 (ACM '78)*. Association for Computing Machinery, New York, NY, USA, pp. 746–752.
- Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*.
- Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, B. H., Van Overveldt, T., Petrou, D., Ramage, D. & Roselander, J. (2019). Towards Federated Learning at Scale: System Design. *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA.
- Bonawitz, Keith, et al., (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*.
- British Science Association (BSA) (2016). One in three believe that the rise of artificial intelligence is a threat to humanity. <https://www.britishtscienceassociation.org/news/rise-of-artificial-intelligence-is-a-threat-to-humanity>
- Brundage, M., Alvin, S. et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, *arXiv: 2004.07213*.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* (3(1)).
- Buschjäger, S. & Pfahler, L. & Buss, J. & Morik, K. & Rhode, W. (2020). On-Site Gamma-Hadron Separation with Deep Learning on FPGAs. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

- 
- Campe, R. & Schneider, M. (1996). *Geschichten der Physiognomik. Text – Bild – Wissen*. Freiburg im Breisgau: Rombach.
- Cardenal, A. , Aguilar-Paredes, C., Galais, C. & Pérez-Montoro, M. (2019). Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure. *The International Journal of Press/Politics* (24 (4)), pp. 1-22.
- Castelvecchi, D. (2016). Can We Open the Black Box of AI? *Nature Magazine*: <https://www.scientificamerican.com/article/can-we-open-the-black-box-of-ai/>
- Chalmers, M., & Galani, A. (2004). Seamful interweaving: heterogeneity in the theory and design of interactive systems. *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pp. 243-252.
- Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). *Trustworthy AI. Reflections on Artificial Intelligence for Humanity*, Springer, Cham, pp. 13-39.
- Chee, F. Y. (2021). EU wants to ban use of AI for surveillance. *Reuters*: <https://www.reuters.com/article/eu-tech-artificialintelligence-idUSL1N2M71DL>
- Chohlas-Wood, A. (2020). Understanding risk assessment instruments in criminal justice. *Brookings*: <https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/>
- Choi, J., Larson, M., Li, X., Li, K., Friedland, G., & Hanjalic, A. (2017). The geo-privacy bonus of popular photo enhancements. *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 84-92.
- Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.
- Chow, R., Jin, H., Knijnenburg, B. & Saldamli, G. (2013). Differential data analysis for recommender systems. *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, pp. 323–326.
- Clancey, W. J. (1983). The epistemology of a rule-based expert system – A framework for explanation, *Artificial Intelligence* 20 (3), pp. 215–251.
- Cremers, A., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A. & Wrobel, S. (2019). *Vertrauenswürdiger Einsatz von Künstlicher Intelligenz*. Fraunhofer Institut für Intelligente Analyse-und Informationssysteme, Sankt Augustin.
- Croy, M.J. (1989). Ethical issues concerning expert systems' applications in education. *AI & Soc* (3), pp. 209–219.
- D’Arcy, J., Gupta, A., Tarafdar, M., & Turel, O. (2014). Reflecting on the “dark side” of information technology use. *Communications of the Association for Information Systems*, 35(1), pp. 5.
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruitingtool-that-showed-bias-against-women-idUSKCN1MK08G>
- De Wit, J. B. F., Das, E. & Vet, R. (2008). What works best: objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology* (27 (1)), pp. 110-115.

- 
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), pp. 554-559.
- Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A. & Quattrociocchi, W. (2017). Mapping social dynamics on Facebook: The Brexit debate. *Social Networks* (50), pp. 6-16.
- Devito, M., Birnholtz, J., Hancock, Jeffery T. & Liu, S. (2018). How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-12.
- Dignum, V. (2017). Responsible Autonomy. *Proceedings of the XXVI International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp. 4698-4704.
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.
- Dove, G., Halskov, K., Forlizzi, J. & Zimmerman, J. (2017). UX design innovation: Challenges for working with machine learning as a design material. *Proc. of 2017 CHI Conference on Human Factors in Computing Systems*, pp. 278-288.
- Druga, S., T.Vu, S., Likhith, E. & Qiu, T. (2019). Inclusive AI literacy for kids around the world. *Fablearn'19*, New York City, USA.
- Dudley, J. J., & Kristensson, P. O. (2018). A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2), pp. 8:1-8:37.
- Dudley, J.J. & Kristensson, P.O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 1, Article 1.
- Dwork, C. (2008). Differential Privacy: A Survey of Results. In: Agrawal M., Du D., Duan Z., Li A. (eds) *Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science (4978)*. Springer, Berlin, Heidelberg.
- Eiband, M., Völkel, S. T., Buschek, D., Cook, S. & Hussmann, H. (2019). When people and algorithms meet: user-reported problems in intelligent everyday applications. *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, pp. 96-106.
- Eitel-Porter, R., Corcoran, M. & Connolly, P. (2021). *Responsible AI From principles to practice*. Accenture:  
<https://www.accenture.com/za-en/insights/artificial-intelligence/responsible-ai-principles-practice>
- Ekstrand, M. D., & Willemsen, M. C. (2016). Behaviorism is Not Enough: Better Recommendations Through Listening to Users. *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 221-224.
- Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K. & Kirlik, A. (2016). First I "like" it, then I hide it: Folk Theories of Social Feeds. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. pp. 2371-2382.
- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K. & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pp. 153-162.
- Eslami, M., Vaccaro, K., Lee, M. K., Elazari, A., Gilbert, E. & Karahalios, K. (2019). User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-14.


- 
- Fast, E. & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. Association for the Advancement of Artificial Intelligence.
- Fernandez, M., & Bellogin, A. (2020). Recommender Systems and Misinformation: The Problem or the Solution?. Proceedings of OHARS'20: Workshop on Online Misinformation- and Harm-Aware Recommender Systems co-located with 14th ACM Conference on Recommender Systems.
- Fernández-Martínez, C. & Fernández, A. (2020). AI and recruiting software: Ethical and legal implications. *Paladyn, Journal of Behavioral Robotics*, (11 (1)), pp. 199-216.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society.
- Fourney, A., Racz, Miklos Z., Ranade, G., Mobius, M, Horvitz, E. (2017), Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, pp. 2071–2074.
- Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter* 15, pp. 1–10.
- Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication* (14 (2)), pp. 265–285.
- Geiger, G. (2021). How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud. *Vice*: <https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud>
- Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M. & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. Proceedings of The 33rd International Conference on Machine Learning, PMLR 48, pp. 201-210 .
- Goodfellow, I., Papernot, Huang, S., Duan, R., Abbeel, P. & Clark, J. (2017). Attacking Machine Learning with Adversarial Examples. <https://openai.com/blog/adversarial-example-research/>
- Graepel, T., Lauter, K. & Naehrig, M. (2012). ML Confidential: Machine Learning on Encrypted Data. In Kwon, T., Lee, M-K. & Kwon, D. (eds) (2012). *Information Security and Cryptology – ICISC 2012*, pp. 1-21.
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (Volume 70). *JMLR.org*, pp. 1321–1330.
- Habib, H., Pearman, S., Wang, J., Zou, Y., Acquisti, A., Cranor, L. F., ... & Schaub, F. (2020, April). "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-12.
- Hamari, J. & Koivisto, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* (45), pp. 191–210.
- Hamilton, K., Karahalios, K., Sandvig, C. & Eslami, M. (2014). A path to understanding the effects of algorithm awareness. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pp. 631–642.
- Hassan, T. (2019). Trust and Trustworthiness in Social Recommender Systems. Companion Proceedings of The 2019 World Wide Web Conference, ACM, New York, NY, USA, 529–532.




- 
- He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* (56), pp. 9–27.
- Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), pp. 993-1012.
- Henley, J. (2021). Dutch government faces collapse over child benefits scandal. *The Guardian*: <https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal>
- Hernandez-Bocanegra, D. & Ziegler, J. (2021). Effects of interactivity and presentation on review-based explanations for recommendations. INTERACT 2021, Proceedings session: HCAI - XAI.
- Hesamifard, E., Takabi, H., Ghasemi, M. & Wright, R. N. (2018). Privacy-preserving machine learning as a service. *Proceedings on Privacy Enhancing Technologies* (3), pp. 123–142 .
- Hesamifard, E., Takabi, H. & Ghasemi, M. (2017). CryptoDL: Deep neural networks over encrypted data, arXiv preprint arXiv:1711.05189.
- Hill, K. (2020). Wrongfully Accused by an Algorithm. *The New York Times*: <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning* (11), pp. 63–91.
- Holton, R., & Boyd, R. (2021). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) Situating artificial intelligence within a socio-technical framework. *Journal of Sociology* (57(2)), pp. 179-195.
- Holzinger, A. (2018). From Machine Learning to Explainable AI. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), IEEE.
- Howard, A. (2020). Are We Trusting AI Too Much?: Examining Human-Robot Interactions in the Real World. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*.
- HubSpot Global AI Survey (2016). Artificial Intelligence Is Here - People Just Don't Realize It. HubSpot: <https://blog.hubspot.com/marketing/artificial-intelligence-is-here>
- Inman, S., & Ribes, D. (2019). "Beautiful Seams". Strategic Revelations and Concealments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-14.
- Jackson, S. L., Krajcik, J., & Soloway, E. (1998). The design of guided learner-adaptable scaffolding in interactive learning environments. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 187-194.
- Jagadeesh, K., Wu, D. J., Birgmeier, J., Boneh, D. & Bejerano, G. (2017). Deriving genomic diagnoses without revealing patient genomes. *Science* (357 (6352)), pp. 692-695.
- Jannach, D., Zanker, M., Felfernig, A. & Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.
- Jesse, M., & Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* (3).
- Jiawei, S., Vargas, D. V. & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* (23 (5)), pp. 828-841.

- 
- Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press.
- Johnson-Laird, P.N. (1980). Mental models in cognitive science. *Cognitive Science* (4 (1)), pp. 71-115.
- Jugovac, M., & Jannach, D. (2017). Interacting with Recommenders—Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems* (7(3)), pp. 10:1-10:46.
- Kahn, K. & Winters, N. (2017). Child-friendly programming interfaces to AI cloud services. *European Conference on Technology Enhanced Learning*, pp. 566-570.
- Kaiser, F. G., Byrka, K., & Hartig, T. (2010). Reviving Campbell's Paradigm for Attitude Research. *Personality and Social Psychology Review*.
- Kaiser, K. & Rauchfleisch, A. (2018). Unite the Right? How YouTube's Recommendation Algorithm Connects The U.S. Far-Right. *Medium*: <https://medium.com/@MediaManipulation/unite-the-right-how-youtubes-recommendation-algorithm-connects-the-u-s-far-right-9f1387ccfabd>
- Karyda, M., Mekler, E. D., & Lucero, A. (2021). Data Agents: Promoting Reflection through Meaningful Representations of Personal Data in Everyday Life. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-11.
- Knobloch-Westerwick, S., Mothes, C. & Polavin, N. (2020). Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research* (47 (1)), pp. 104-124.
- Kocielnik, R., Avrahami, D., Marlow, J., Lu, D. & Hsieh, G. (2018a). Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 881–894.
- Kocielnik, R., Xiao, L., Avrahami, D. & Hsieh, G. (2018b). Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings ACM Interact. Mob. Wearable Ubiquitous Technology* (2 (2)), pp. 70 - 70:26.
- Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development*, Englewood Cliffs, New Jersey: Prentice-Hall.
- Konstan, J. A., & Riedl, J. (2012a). Recommended for you. *IEEE Spectrum*, 49(10), pp. 54–61.
- Konstan, J.A. & Riedl, J. (2012b). Recommender systems: from algorithms to user experience. *User Model User-Adap Inter* (22), pp. 101–123.
- Koroleva, K. & Novak, J. (2020). How to Engage with Sustainability Issues We Rarely Experience? A Gamification Model for Collective Awareness Platforms in Water-Related Sustainability, *Sustainability* (12 (2)), pp. 712.
- Koroleva, K., Krasnova, H. & Günther, O. (2010). 'STOP SPAMMING ME!' - Exploring Information Overload on Facebook. *AMCIS 2010 Proceedings*, pp. 447.
- Koroleva, K., Melenhorst, M., Novak, J., Gonzalez, S.L.H., Fraternali, P. & Rizzoli, A.E. (2019). Designing an integrated socio-technical behaviour change system for energy saving. *Energy Informatics* (2 (30)).
- Kou, Y. & Gui, X. (2020). Mediating Community AI Interaction through Situated Explanation: The Case of AI Led Moderation. *Proceedings of the ACM on Human-Computer Interaction* (4), pp. 102:1 -102:27.

- 
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I. & Wong, W.K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *IEEE Symposium on Visual Languages and Human Centric Computing*, pp. 3-10.
- Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.
- Lambrech, A., & Tucker, C. E. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*.
- Langer, M., Oster, D., Speith, T., Hermanns, H., Ka"stner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*.
- Larson, M., Zito, A., Loni, B. & Cremonesi, P. (2017). Towards Minimal Necessary Data: The Case for Analyzing Training Data Requirements of Recommender Algorithms. *Proceedings of the RecSys 2017, Workshop on Fairness, Accountability and Transparency in Recommender Systems*.
- Lecher, C. (2018). What happens when an algorithm cuts your health care. *The Verge*: <https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy>
- LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature* (521), pp. 436–444.
- Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A. & Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1-35.
- Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*.
- Li, C-Y. (2017). Why do online consumers experience information overload? An extension of communication theory. *Journal of Information Science* (43 (6)), pp. 835-851.
- Lin, J., Liu, B., Sadeh, N., & Hong, J. I. (2014). Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. *10th Symposium On Usable Privacy and Security*, pp. 199-212.
- Lindell, Y. (2020). Secure multiparty computation. *Communications of the ACM* (64 (1)).
- Lindgren, S., & Holmström, J. (2020). A social science perspective on artificial intelligence: building blocks for a research agenda. *Journal of digital social research*, 2(3), pp. 1-15.
- Liu, Z., Zhao, Z. & Larson, M. (2019). Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval. *Proceedings of the 2019 International Conference on Multimedia Retrieval*.
- Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-16.
- Lukowicz, P. (2020) European Project Humane AI: Human-Centered Artificial intelligence: <https://www.humane-ai.eu/>
- Lusted, L. B. (1978). General problems in medical decision making with comments on ROC analysis. *Seminars in nuclear medicine*, WB Saunders (8(4)), pp. 299-306.


- 
- Marin, L. & Roeser, S. (2020). Emotions and Digital Well-Being. The Rationalistic Bias of Social Media Design in Online Deliberations. In Approach, C. Burr and L. Floridi (eds.) (2020). *Ethics of Digital Well-being: A Multidisciplinary*, Springer.
- Marin, L. (2020). Three contextual dimensions of information on social media: lessons learned from the COVID-19 infodemic. *Ethics and information technology*, pp. 1–8.
- Marsili, N. (2020). Retweeting: its linguistic and epistemic value. *Synthese*.
- McCarthy, J. (1979). *Ascribing Mental Qualities to Machines*. Stanford University California, Department of Computer Science.
- Michael, L. & Otterbacher, J. (2014). Write Like I Write: Herding in the Language of Online Reviews. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM*, pp. 356-365.
- Miller, T., Howe, P. & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.
- Mohallick, I., De Moor, K., Özgöbek, Ö., & Gulla, J. A. (2018). Towards new privacy regulations in europe: Users' privacy perception in recommender systems. In *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*, Springer, Cham, pp. 319-330 .
- Mohassel, P. & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. *2017 IEEE Symposium on Security and Privacy*.
- Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574-2582.
- Morris, T. H. (2019). *Experiential learning – a systematic review and revision of Kolb's model*. Interactive Learning Environments.
- Moyer-Gusé, E. (2008). Toward a Theory of Entertainment Persuasion: Explaining the Persuasive Effects of Entertainment-Education Messages. *Communication Theory* (18 (3)), pp. 407–425.
- Naul, E. & Liu, M. (2019). Why Story Matters: A Review of Narrative in Serious Games. *Journal of Educational Computing Research*.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* (2 (2)), pp. 175-220.
- Norman, D. (1983). Some Observations on Mental Models. In D. Gentner & A. Stevens (Eds.), *Mental Models*. Psychology Press, pp. 7–14.
- Novak, J. & Peranovic, P. (2004). Supporting Experiential Learning through Online/Onsite Interaction and Collaborative Use of Mobile Devices. *Workshop on Interaction Design for CSCL in Ubiquitous Environments at Mobile HCI - 6th International Conference for Mobile Human-Computer Interaction*.
- Novak, J., Melenhorst M., Micheel, I., Pasini, C., Fraternali, P. & Rizzoli, A.E. (2018). Integrating behavioural change and gamified incentive modelling for stimulating water saving. *Environmental Modelling and Software* (102), pp. 120-137.
- Novak, J., Wieneke, L., Düring, M., Micheel, I., Melenhorst, M., Morón, J.G., Pasini, C., Tagliasacchi, M. & Fraternali, P. (2014). histoGraph – A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections. *Proceedings of 18th International Conference Information Visualisation* (4).

- 
- Nyhan, B. & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* (32), pp. 303–330.
- O’Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2014). Down the (White) Rabbit Hole. *Social Science Computer Review* (33(4)), pp. 459–478.
- Obrenović, Ž. (2012). Rethinking HCI education: teaching interactive computing concepts based on the experiential learning paradigm. *Interaction* (19(3)), pp. 66-70.
- One Hundred Year Study on Artificial Intelligence (AI100) (2016). Stanford University: <https://ai100.stanford.edu>.
- Osiurak, F., Navarro, J., & Reynaud, E. (2018). How our cognition shapes and is shaped by technology: a common framework for understanding human tool-use interactions in the past, present, and future. *Frontiers in psychology* (9), pp. 293.
- Pană, L. (1973). Elements of Artificial Ethics for Cognitive and Moral Agents. *Noesis*, 33, 39.
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506-519.
- Pariser, E. (2012). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
- Pfahler, L. & Morik, K. (2020): Fighting Filter Bubbles with Adversarial Training. 2nd Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia.
- Ploderer, B., Reitberger, W., Oinas-Kukkonen, H. & van Gemert-Pijnen, J. (2014). Social interaction and reflection for behaviour change. *Pers Ubiquit Comput* (18), pp. 1667–1676.
- Potthast M., Köpsel S., Stein B. & Hagen M. (2016) Clickbait Detection. In: Ferro N. et al. (eds) *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science* (9626), Springer, Cham.
- Quattrociocchi, W. & Scala, A. & Sunstein, C. R. (2016). Echo Chambers on Facebook.
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D. & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The journal of the learning sciences* (13 (3)), pp. 337-386.
- Raafat, R. M., Chater, N. & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences* (13(10)), pp. 420-428.
- Radha, M., Willemsen, M. C., Boerhof, M., & IJsselsteijn, W. A. (2016). Lifestyle Recommendations for Hypertension Through Rasch-based Feasibility Modeling. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pp. 239–247.
- Raghavan, M., Barocas, S., Kleinberg, J. & Levy K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT\* '20)*, pp. 469–481.
- Rao, A., Palaci, F. & Chow, W. (2019). *A practical guide to Responsible Artificial Intelligence (AI)*. PwC: [www.pwc.com/rai](http://www.pwc.com/rai)
- Resnick, M., Berg R. & Eisenberg, M. (2000) Beyond black boxes: bringing transparency and aesthetics back to scientific investigation. *Journal of the Learning Sciences* (9 (1)), pp. 7–30.
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F. & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and*

- 
- Transparency (FAT\* '20). Association for Computing Machinery, New York, NY, USA, pp. 131-141.
- Richardson, R., Schultz, J. & Sutherland, V. (2019). Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems. AI Now Institute.
- Rini, R. (2017). Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal* (27 (2S)).
- Rudin, C. & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*.
- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* (1), pp. 206-215.
- Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2019). Beyond Behavior: The Coach's Perspective on Technology in Health Coaching. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-14.
- Ryffel, F. A. & Wirth, W. (2020). How perceived processing fluency influences the illusion of knowing in learning from TV reports. *Journal of Media Psychology: Theories, Methods, and Applications* (32 (1)), pp. 2-13.
- Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N. & Beard, N. (2019). Integrating Ethics within Machine Learning Courses. *Association for Computing Machinery* (19 (4)), pp. 1-26.
- Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P. A., & Santos, I. (2019). Can i opt out yet? gdpr and the global illusion of cookie control. *Proceedings of the 2019 ACM Asia conference on computer and communications security*, pp. 340-351.
- Schäfer, H., & Willemsen, M. C. (2019). Rasch-based tailored goals for nutrition assistance systems. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 18-29.
- Schmitt, J. B., Debbelt, C. A. & Schneider, F. M. (2018). Too much information? Predictors of information overload in the context of online news exposure. *Information, Communication & Society* (21 (8)), pp. 1151-1167.
- Schnabel, T., Amershi, S., Bennett, P. N., Bailey, P. & Joachims, T. (2020). The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, pp. 991-1000.
- Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE TVCG* (16 (6)), pp. 1139-1148.
- Serenko, A., & Turel, O. (2015). Integrating technology addiction and use: An empirical investigation of Facebook users. *AIS Transactions on Replication Research* (1 (1)), pp. 2.
- Sharma, P., & Hannafin, M. J. (2007). Scaffolding in technology-enhanced learning environments. *Interactive learning environments* (15 (1)), pp. 27-46.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* (36 (6)), pp. 495-504.
- Shneiderman, B. (2021). Human-Centered AI. *Issues in Science and Technology* (37 (2)), pp. 56-61.



- 
- Slater, M. D., & Rouner, D. (2002). Entertainment—education and elaboration likelihood: Understanding the processing of narrative persuasion. *Communication theory* (12 (2)), pp. 173-191.
- Sokol, K. & Flach, P. (2018). Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pp. 5868-5870.
- Starke, A. D., Willemsen, M. C., & Snijders, C. (2017). Effective User Interface Designs to Increase Energy-efficient Behavior in a Rasch-based Energy Recommender System. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pp. 65–73.
- Starke, A. D., Willemsen, M. C., & Snijders, C. (2020). With a little help from my peers: Depicting social norms in a recommender interface to promote energy conservation. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 568–578.
- Strickland, E. (2019). Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients. *IEEE Spectrum*: <https://spectrum.ieee.org/the-human-os/biomedical/ethics/racial-bias-found-in-algorithms-that-determine-health-care-for-millions-of-patients>
- Strömbäck, J. (2005). In search of a standard: Four models of democracy and their normative implications for journalism. *Journalism studies* (6(3)), pp. 331-345.
- Su, X., Özgöbek, Ö., Gulla, J. A., Ingvaldsen, J. E., & Fidjestøl, A. D. (2016). Interactive mobile news recommender system: A preliminary study of usability factors. In *2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, IEEE, pp. 71-76.
- Szolovits, P. & Pauker, S. G. (1979). Computers and clinical decision making: Whether, how, and for whom? *Proceedings of the IEEE* (67 (9)), pp. 1224-1226.
- Telefónica (2018). AI Principles of Telefónica. <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>
- Thornhill, J. (2020). Trusting AI too much can turn out to be fatal. *Financial Times*: <https://www.ft.com/content/0e086832-5c5c-11ea-8033-fa40a0d65a98>
- Trattner, C., & Elswiler, D. (2017). Investigating the Healthiness of Internet-Sourced Recipes: Implications for Meal Planning and Recommender Systems. *Proceedings of the 26th International Conference on World Wide Web*, pp. 489–498.
- Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019, November). (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 973-990.
- van Koningsbruggen, G.M. & Das, E. (2009). Don't derogate this message! Self-affirmation promotes online type 2 diabetes risk test taking. *Psychology and Health* (24 (6)), pp. 635-649.
- Vehof, H., Heerdink, E., Sanders, J. & Das, E. (2019). Associations between characteristics of online diabetes news and readers' sentiment: Observational study in the Netherlands. *Journal of Medical Internet Research* (21 (11)).
- Versenyi, L. (1974). Can robots be moral? *Ethics* (84 (3)), pp. 248-259.
- Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., & Helberger, N. (2020). Recommenders with a mission: assessing diversity in news recommendations. *arXiv preprint arXiv:2012.10185*.

- 
- Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-6.
- Wang, D., Yang, Q., Abdul, A. & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA (601), pp. 1-15.
- Wang, F. & Rudin, C. (2015). Falling rule lists. *Artificial Intelligence and Statistics*. PMLR.
- Wang, Y., Si, C. & Wu, X. (2015). Regression model fitting under differential privacy and model inversion attack. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- Weiser, M. (1994). Creating the invisible interface. *Proceedings of the 7th annual ACM symposium on User interface software and technology (UIST '94)*.
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co.
- Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kazianus, E., Mathur, V., West, S. M., Richardson, R., Schultz, J. & Schwartz, O. (2018). *AI Now Report 2018*, AI Now Institute, New York University.
- Winter, M. & Jackson, P. (2020). Flatpack ML: How to Support Designers in Creating a New Generation of Customizable Machine Learning Applications. *Proceedings of HCI 2020*, Springer LNCS, Volume 12201, pp. 175-193.
- Wohn, D. Y. & Bowe, B. J. (2016). Micro Agenda Setters: The Effect of Social Media on Young Adults' Exposure to and Attitude Toward News. *Social Media + Society* (2 (1)).
- Wollebæk, D., Karlsen, R., Steen-Johnsen, K. & Enjolras, B. (2019). Anger, Fear, and Echo Chambers: The Emotional Basis for Online Behavior. *Social Media + Society* (5 (2)), pp. 1-14.
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1-13.
- Zhang, Q., Wu, Y. N. & Zhu, S. (2018). Interpretable Convolutional Neural Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8827-8836.
- Zhang, Y., Bellamy, R. & Varshney, K. (2020). Joint Optimization of AI Fairness and Utility: A Human-Centered Approach. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, pp. 400-406.
- Zimmermann, M. R. (2018). *Teaching AI: exploring new frontiers for learning*. International Society for Technology in Education.