Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

How to design human-centered AI solutions – review of the human-centered AI design

10.8.2020
Ala-Luopa, Saara; Lehtiö, Anu; Olsson, Thomas; Väänänen, Kaisa

RESEARCH QUESTIONS:

1. What are the best HCD methods and practices in AI applications?
2. What are the quality attributes in HCD of AI applications?
3. What makes good UX in human-centered design of AI applications?

Table of Contents

1. What is human-centered AI?

There is no one definition for human-centered artificial intelligence. However, versatile definitions include commonalities. Stanford University, UC Berkeley, and MIT have established human-centered AI (HAI) research institutes. Their HAI

research strategies emphasize the humanistic and ethical viewpoint of the next frontier of AI: AI is to *enhance* humans rather than replace them (Xu 2019). Also, HumaneAI (Crowley et al. 2019) research community support this thought: The project aims to "develop paradigms that allow humans and AI systems including service robots and smart environments to interact and collaborate in a way that *enhances* human abilities and empowers people." According to European Commission Ethics Guidelines (2019), AI systems should *empower* human beings, allowing them to make informed decisions and foster their fundamental rights. This remark is notified also in the industry: AI technology should be used to *enhance* and augment human potential rather than replace it (e.g. Wärnestål 2019). However, Google provides recommendations in their People + AI Guidebook whether to automate or augments tasks at hand: recommendation is to automate tasks that are difficult or unpleasant and ideally ones that can be agreed to have the "correct" way to do it. Bigger processes, that people enjoy doing or carry social value should be augmented, meaning AI rather complements existing human abilities and give them "superpowers" instead of automating a task away entirely (Google 2019). In summary, 'enhance' seems to be suitable word to describe appropriate AI-human collaboration.

What is notable in AI systems, is their increasing agency and learning capabilities: the potential of AI technologies makes possible services and systems that act on our behalf and take their own initiative. According to MIT, the design, development, and deployment of AI systems that learn from and collaborate with humans in a meaningful way is defined by two goals: the AI system must continually *improve* by learning from humans (1), while creating an effective and fulfilling human-robot interaction experience (2). Xu (2019) states, that with the addition of learning capabilities in AI-based machine intelligence, human-machine relationships have shifted from human-computer interaction to human-machine *integration* and human-machine *teaming*. Wärnestål (2019) even claims, that we are transitioning from designing tools to designing partners (as a comparison to designing moment-to-moment tools). Despite the specific role of AI solution e.g. as a "partner", the capability to learn and improve system performance is one of the most essential features of AI solution.

1.1. "Human-in-the-loop" philosophy

A common concern in the increasing agency of AI systems is the loss of human control. To avoid this risk, human-centered AI calls for proper oversight mechanisms which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches. "Human-in-the-loop philosophy" focuses on creating workflows where an AI learns from the human operator while intuitively making the human's work more efficient. (Hulkko 2018.) This is also one of the research streams in HumaneAI (Crowley et al. 2019), where Human-in-the-Loop Machine Learning, reasoning, and planning is defined as following: "Allowing humans to not just understand and follow the learning, reasoning, and planning process of AI systems (being explainable and accountable), but also to seamlessly interact with it, guide it, and enrich it with uniquely human capabilities, knowledge about the world, and the specific user's personal perspective." Thus, users should be willing to continuously tech, supervise and guide AI system. While we expect AI system to *enhance* our daily

or professional lives, we should be ready to help the system to fulfil our expectations (also users of AI systems need to put effort to human-AI collaboration).

1.2. Human judgement

Microsoft's Guidelines for AI-Human Interaction emphasize user understanding and control, while addressing ways for the system to "make clear why the system did what it did" (understandability) and "learn from user behavior" (adaptability). When an AI model makes a mistake, we need human judgment: humans need to gauge the context in which an algorithm operates and understand the implications of the outcomes (Deloitte 2019). HCI design is required to ensure that human operators can quickly and effectively take over the control of an intelligent system in an emergency, so that fatal accidents such as the accidents of autonomous cars can be avoided (Xu 2019.) The challenge is how can we spot the "machine-made mistake", how we supervise the system (in a way that system still "enhances" our lives without being too time-consuming) and do we have the knowledge to teach the system (who should and could teach the system, and what data this this knowledge is based on?

1.3. Context awareness of intelligent systems

Previous research has studied human interaction with intelligent context-aware computing systems, e.g. how to design for understandability and control of the underlying sensing systems (see Amershi: 3, 23). However, AI is a probabilistic system: instead of being programmed, it needs to be taught. Firstly, AI solutions must have context to be able to utilize their learning (IBM Design for AI 2019). Secondly, they must share an understanding of a problem's larger context to properly cooperate in developing a solution. (Crowley et al 2019). IBM's AI/Human Context Model describes understanding as "what AI needs to know, the process of putting data in domain context." Understanding is followed by reasoning, which refers to the system's logic to decide for the best courses of action. This sums up as knowledge, which refers to all past data, insights and attributes, and expression to the system response to the user. Of course, user's reaction to the system's expressions gain machine knowledge, and thus the user teaches the system to improve. Crowley et al. (2019) describe this as Multimodal Perception and Modelling: Enabling AI systems to perceive and interpret complex real-world environments, human actions, and interactions situated in such environments and the related emotions, motivations, and social structures. This requires enabling AI systems to build up and maintain comprehensive models that in their scope and level of sophistication, should strive for more human-like world understanding and include common sense knowledge that captures causality and is grounded in physical reality.

1.4. Useful and usable AI

Instead of being force-fit to accommodate technical capabilities or requirements, human-centered AI must primarily address *user needs and values* and it must have a clear purpose: What is the problem or the need that should be solved, and is AI system a right solution for it? AI should be designed to align with the norms and

values of the user group in mind (Google 2019, IBM 2019). Recommended actions to take are e.g. to consider the culture that establishes the value systems that are designing within and aim for understanding the user values (IBM Everyday Ethics for AI 2019).

Purpose refers to the reason for the user to engage with the system, and this will evolve as the user and system grow with each other (IBM 2019). AI needs to be useful and usable: it can provide the functions required to satisfy target users' needs in the valid usage scenarios of their work and life, and it should be easy to learn and use via optimal UX created by effective HCI design. HCI professionals can identify usage scenarios based on HCI methods such as ethnographic studies and contextual inquiries, and helping mine user needs, behavioral patterns, and usage scenarios (Xu 2019.) It is important to notify, what kind of solution is usable to who and when, and in which context and design suitable solutions for specific user needs.

2. Design (and evaluation) methods for HCAI

2.1. Setting expectations

Setting realistic user expectations is important in human-centered AI. Mental models help set expectations about product capabilities and its expected value. Introducing AI solution includes different stages and mental models form and change along the way (Google 2019.) Amershi et al. (2019) guidelines 1 & 2 focus on setting the right expectations during initial interactions between a person and an AI system: it is important to make clear *what the system can* do and *how well* the system can do what it can do. Managing expectations is recommended to avoid misleading or frustrating users during interaction with unpredictable adaptive agents (see Amershi 16, 20, 33).

However, explaining specific product capabilities while providing a high-level mental model of the AI solution is tricky to balance. Recommendation is to *explain the benefit* instead of technology but offer more detailed information if need arises (for explainability and transparency). It is not recommended to introduce too much AI-driven features at once: onboarding, setting up the interaction relationship between the user and the product, is good to keep short since users learn over time and they do not need to know everything at once. Expectations are remarkably important in the adoption phase, since trust in the app's capabilities depends on the expectations for how the system should work and how alerts are worded (Google 2019.)

It is notable, that anthropomorphic or human-like products might set unrealistic expectations. Therefore, disclosing the algorithm-powered nature of these kinds of interfaces is a critical onboarding step and a subject of ongoing research (Google 2019).

2.2. Interaction

Current HCI methods were originally created for non-intelligent solutions, and thus traditional approach is not suitable when designing AI systems. Personalization and adaptation will play an even greater role in the design and implementation of human-

centered AI, because the behavior of intelligent systems develops over time. AI system can adapt to different needs, skills, and abilities of individual users. Relationships can be life-long, which ascertains on the potential of an even higher degree of inclusive design for all (Xu 2019, Wärnestål 2019.) The user's relationship with an AI system can evolve through back-and-forth interactions that reveal their strengths and weaknesses (Google 2019).

Amershi et al. (2019) presents guidelines for human-AI interaction over time in a relation to learning abilities of the system. AI system should remember recent interaction, learn from user's behavior and update and adapt cautiously without disruptive changes. Encouraging user to provide feedback can have immediate consequences, or it can affect to system behavior in the future. It is recommended to let the user customize the system while users should be informed when AI system adds or updates its capabilities. During interaction, AI system should provide time services based on context (timely guidance) and show contextually relevant information, matching relevant social norms and mitigating social biases (experience is as expected and free from undesirable and unfair stereotypes and biases). Possible area for developing autonomous intelligent systems capable of following social and moral norms is to *identify the norms of the specific community* in which the autonomous systems are to be deployed and, in particular, *norms relevant to the kind of tasks* that the autonomous systems are designed to perform (IEEE, 2017).

Major driver of the final user experience is designing and evaluating a reward function, also called "objective function", or "loss function." This determines the action or behavior the system will try to optimize for. Key decisions are 'binary classifiers' (false positives and false negatives), and precision and recall. To evaluate the reward function, it is recommended to assess inclusivity, monitor the user experience and metrics and imagine potential pitfalls (Google 2019.)

2.3. Feedback & control

As AI system learn over time, user experience may change over time: collecting implicit feedback (actions while using the product) and explicit feedback (given feedback to improve the product) is important part of user experience in AI systems. Feedback and control mechanisms are critical to improving your underlying AI model's output and user experience. It is important to understand when people want to maintain control, and to help user control the aspects of the experience they want to, as well as easily opt out of giving feedback. (Google 2019.) When users have the right level of control over the system, they're more likely to trust it. Especially explicit feedback can help the user feel more in control of the product. Feedback mechanisms partner closely with mental models and explainability and how to tune AI (Google 2019.)

Communication is important: Users should know, what information is being collected and why, and how feedback will change their experience or benefit them. In addition, it is important to understand, why people give feedback (e.g. material rewards) (Google 2019.)

2.4. Metrics for success / evaluation metrics

In addition to user needs, metrics for success should be identified (Google 2019). Human-centered evaluation metrics are particularly important when a model is employed to assist humans as in decision-making tasks or mixed-initiative systems. AI designers should go beyond aggregate, single-score performance numbers when evaluating the capabilities and limitations of an AI model. They should use multiple and realistic benchmarks for evaluation and include human-centered evaluation metrics when examining the behavior and performance of an AI. These metrics could e.g. performance explainability: can the human anticipate ahead of time when the system will make a mistake? Performance explainability makes a model more human-centered because it enables people to better understand and anticipate when the model might make mistakes so that the human can take over when needed. Fairness: does the model have comparable performance on different demographic groups? Does the system allocate a comparable amount of resources to such subgroups? Interpretability: how well might a human understand how the model ends up to a certain decision? Discussions around these metrics have led to several open source contributions in the form of libraries for computing and sometimes optimizing for such human-centered metrics: InterpretML, FairLearn, AI Explainability 360 (Nushi 2019.)

2.5. When things go wrong

According to People + AI Guidebook by Google (2019), AI errors are also opportunities: they can support faster learning by experimentation, help establish correct mental models, and encourage users to provide feedback. To define error is important, because it is deeply connected to the expectations of the AI system. In addition to system and user errors, AI solution may include context errors which are based on the system's assumptions about the user.

According to Amershi et al. (2019), AI systems should support efficient invocation, dismissal and correction: system should be easy to request when needed, and it should be easily ignored if unwanted. Also, it should be easy to edit and refine. AI system services should be easily scoped when in doubt and users should have access to explanation, why system did what it did.

The inherent complexity of AI-powered systems can make identifying the source of an error challenging. It is important to discuss inside the design team how errors are discovered, and sources discerned. The trick isn't to avoid failure, but to find it and make it just as user-centered as the rest of the product: the product needs to provide ways for the user to continue their task and to help the AI improve (Google 2019).

3. Ethical viewpoints

3.1. Data & fairness (AI developers & designers)

Data collection and data evaluation is essential in designing AI solutions. At first, designer should define what data is required and what is the data source. Translating user needs to data needs means determining the type of data to training the AI model.

Predictive power, relevance, fairness (=data quality), privacy, and security should be considered in ML models. (Google 2019, IBM 2019.) AI must be designed to protect user data and preserve the user's power over access and uses. Users should always maintain control over what data is being used and in what context, and the data should be protected. Users should be allowed to deny service or data by having the AI ask for permission before an interaction or providing the option during an interaction. (IBM Everyday Ethics for AI 2019.)

Both Google and IBM stress the uncertain nature of data: Bias can be introduced into the ML model in every stage of the development, and there is no such thing as truly neutral data. *Machine learning by its very nature is always a form of statistical discrimination* (Google 2019, IBM 2019.) Addressing bias requires an understanding of the underlying structural inequalities (Whittaker et al., 2018; United Nations, 2018). Whether explicit or implicit, biases are the symptom of a lack of diversity within the people who build the technology (Li, 2018). Unfair bias could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle (European Commission Guidelines for Trustworthy AI 2019). Diverse teams help to represent a wider variation of experiences to minimize bias (IBM Everyday Ethics for AI 2019.), and a way of mitigating bias is aimed at the trainers and creators of the AI. By making them aware of their own prejudices, we have a better chance of keeping it out of the algorithms (Kwan 2018).

According to Saif & Ammanath (2020), fairness is related to trustworthy AI: AI must be designed and trained to follow a fair, consistent process and make fair decisions. To avoid problems related to fairness and bias, companies first need to determine *what constitutes "fair."* Companies also need to actively look for bias within their algorithms and data, making the necessary adjustments and implementing controls to help ensure additional bias does not pop up unexpectedly. When bias is detected, it needs to be understood and then mitigated through established processes for resolving the problem and rebuilding customer trust.

Evaluation and collecting the data, data sourcing, accurate data labels and rater tools are crucial. Testing and tuning the model is an ongoing process. AI designers need to have the right data and enough of it: quantity and quality matters. Data sampling helps ensure that dataset matches the real world and is representative of the whole. Data completeness indicates whether all the data that is needed is available in the data resources and that data doesn't have gaps. Consolidating data means making data to work together. Data should be consistent and rich: "say the same thing" and give the essence of what's really going on (Google 2019, IBM 2019.) As a society we still need to decide what data should be allowed for algorithms to use to make inferences (Matsakis, 2018).

In order to protect against biases in algorithmic decision-making, companies must conduct periodic audits to ensure "algorithmic hygiene" before, during, and after implementing AI tools. For example: 1. conducting audits: to frequently examine algorithms for biases and to delete any biased associations, 2. get feedback from users: In addition to internally gauging algorithms' performance, it's vital to seek out feedback from the customers as well, 3. ensuring that AI tools are transparent and explainable. For any given automated behavior, *developers must be able to explain*

why the algorithm engaged in that particular course of action. (Ramamoorthy 2019). According to Google (responsible AI 2019), recommended practices for fairness are: 1. Design the model using concrete goals for fairness and inclusion, 2. use representative datasets to train and test your model, 3. check the system for unfair biases and 4. analyze performance.

3.2. Explainable AI (end-users)

Explainable AI (XAI) enables users to understand the algorithm and parameters used, which is intended to address the AI black-box problem. From an HCI perspective, there is no guarantee that the target users of an XAI system will be able to understand it. Explanation creates understanding – it helps users to understand the systems decision process and helps to create realistic mental models of the system's capabilities and limits (Google People + AI Guidebook 2019, IBM Everyday Ethics for AI 2019). The goal of XAI should be to ensure that target users can understand the outputs, thus helping them improve their decision-making efficiency (Xu 2019.) *User should be able to ask why an AI is doing what it is doing.* Also, decision-making processes must be reviewable, especially if the AI is working with highly sensitive personal information data. AI must be able to provide a sufficient explanation of recommendations, the data used, and the reasoning behind the recommendations, and designing teams should have and maintain access to a record of an AI systems decision processes. However, there are situations where users may not have access to the full decision process that an AI might go through, e.g., financial investment algorithms. Thus, it should be ensured that an AI systems' level of transparency is clear. Users should stay generally informed on the AI systems intent even when they can't access an analysis of the AI process (IBM Everyday Ethics for AI 2019.)

According to Google People + AI Guidebook (2019), explanations help users to evaluate AI systems. *Optimizing for understanding* is recommended instead of completely explaining the system: in some cases, there may be no explicit, comprehensive explanation for the output of a complex algorithm, or it may be too complex to explain. Algorithms should offer people "counterfactual explanations", or disclosure about the decision and provide the smallest change that can be made to obtain a desirable outcome (Wachter et al. 2018). For example, an algorithm that calculates loan approvals should explain not only why credit was denied, but also what can be done to reverse the decision: *users don't necessarily need to understand how a machine learning system works to know why it reached a certain decision.* (Matsakis, 2018). Managing influence on user decisions is important, because AI systems often generate output that the user needs to act on. Confidence level (statistical measure of how certain a prediction or outcome is) can be critical in informing the users decision making and calibrating their trust (Google People + AI Guidebook 2019).

Fundamental design decision in explainable AI begins with 1) understanding the data that is being used to train AI, 2) choosing the proper type of decision engine and, 3) selecting algorithms that explain decisions after they are made. These three steps are often referred to as data explainability, model explainability, and post hoc explainability. AI community has mainly focused on post hoc explainability (Reese / GigaOm) Large body of work exists and continues to grow around how to increase

transparency or explain the behaviors of AI systems (See Amershi et al. 14, 21, 23, 36, 38, 44].

It is noteworthy, that 'explainability' is shown to have different meanings, and the needs vary considerably according to the audience: Designers, developers, users or affected people do not need the same level and type of explanation. (European Parliament 2019: Understanding algorithmic decision-making: Opportunities and challenges.)

3.3. Transparency (AI developers & designers)

Transparency in AI means using intuitive language to talk about the AI systems under a development, how they work and what they are capable of. In addition, transparency means explaining where the data comes from. (MindAI 2018.) Transparency creates explainability. There are lots of open questions regarding *what constitutes a fair explanation and what level of transparency is sufficient.* It's necessary to ask: Transparent to whom and for what purpose? (Matsakis 2018). Too much transparency as letting people know how decisions are made can allow them to "game" the system and orient their data to be viewed favorably by the algorithm (Gillespie 2016). Thus, AI holds a "transparency paradox": while generating more information about AI might create real benefits, it may also create new risks. Explanations can be hacked and releasing additional information may make AI more vulnerable to attacks, and disclosures can make companies more susceptible to lawsuits or regulatory action. To navigate this paradox, *organizations will need to think carefully about how they're managing the risks of AI*, the information they're generating about these risks, and how that information is shared and protected. (Burt 2019.) Many software companies choose not to disclose what algorithms they use or what data they use to train them, which is often reasoned as to protect intellectual property or prevent a security breach. (Hume 2018.)

According to European Parliament 'Understanding algorithmic decision-making: Opportunities and challenges' (2019) report, transparency should not be the ultimate solution for users or people affected by the algorithmic decisions since source code is illegible to non-experts. Transparency mainly benefits e.g. independent experts, NGOs, evaluation bodies or data protection authorities (DPA). However, while automation has the potential to make us more human by taking off the tedious and repetitive tasks humans are not good at, it will require us to be more critical and reflect on our practice to find where our human intelligence will be necessary (Hume, 2018). Individual decision-making will be improved when people know they are interacting with AI systems. For example, anthropomorphism might create problems in AI transparency: customer service calls, website chatbots, and interactions on social media and in virtual reality may become progressively less evidently artificial. When users know they are interacting with AI system, they can make judgements about the advantages and limitations of the system and then choose whether to work with it or seek human help. AI transparency may help humans direct their sincerity primarily toward people, not robots. (Engler 2020.) Another critical aspect to AI research is how individuals are impacted by being part of the algorithmic decision-making process with non-human actors in the decision (Martin 2018).

According to Deloitte (2019), *the level of transparency depends on the impact of the technology*. The more impact an advanced or AI-powered algorithm has, the more important it is that it is explainable and that all ethical considerations are in place. Creating transparent AI requires firstly, technical steps and technical correctness: the developer of the model must be able to explain how they approached the problem, why a certain technology was used, and what data sets where used. Secondly, consideration whether the outcomes of the model are statistically sound: whether certain groups are under-represented in the outcomes. This step can help to detect hidden biases in data because only humans, who understand the context in which the data has been collected, can spot possible biases in the outcome of the model. Thirdly, AI models should be validated to enable organizations to understand what is happening in the model and to make the results explainable.

3.4. Accountable AI (end-users & AI developers & designers)

Accountability is the most important requirement as far as the protection of individuals is concerned. Governance and accountability issues refer to those who create the ethics standards for AI, who governs the AI system and data, who maintains the internal controls over the data and who is accountable when unethical practices are identified. (Mintz 2019.) Recommended actions are e.g. to make company policies clear and accessible to design and development teams since from the beginning so that no one is confused about issues of responsibility or accountability, and keep detailed records of the design processes and decision making (IBM Everyday Ethics for AI 2019.) The internal auditors should assess risk, determine compliance with regulations and report their findings directly to the audit committee of the board of directors. Corporate governance is essential to develop and enforce policies, procedures and standards in AI systems. Chief ethics and compliance officers have an important role to play, including identifying ethical risks, managing those risks and ensuring compliance with standards. (Mintz 2019.)

Auditability (which enables the assessment of algorithms), data and design processes play a key role especially in critical applications. Auditing is the function of examining data to determine whether it is accurate and reliable, and that the system used to generate it is operating as intended (Mintz 2019). Oversight agencies and supervisory authorities should play a central role. It is critical that they have all the means necessary to carry out their tasks. Accountability can be achieved via complementary means such as algorithmic impact assessments (AIA's), auditing and certification. If appropriate accountability measures are taken, in certain situations *algorithmic decision-systems have the potential to improve transparency and reduce unfairness and discrimination.* (European Commission Guidelines for Trustworthy AI 2019.)

According to Kwan (2018), the best way to keep track of accountability is to keep accurate and detailed records of the AI decision-making processes: the processes and data by which the decisions come should be transparent, so that if anything goes wrong, some third-party auditor is able to retrace the steps leading up to the outcome to locate the source of the problem.

3.5. Responsible AI (AI developers & designers)

Responsible AI means that AI system has all the ethical considerations in place, and it is aligned with the core principles of the developing company. Responsible AI helps organizations to regain control over the AI models that are deployed. (Deloitte 2019.) Assigning responsibility for AI governance is essential, because without clear responsibilities, no one is accountable. (Sondergaard 2019.) Responsibility creates accountability.

Google (responsible AI practices 2019) states that reliable, effective user-centered AI systems should be designed following general best practices for software systems, together with practices that address considerations unique to machine learning. Their top recommendations are following:

1. Use a human-centered design approach: The way actual users experience the system is essential to assessing the true impact of its predictions, recommendations and decisions. This means e.g. designing features with appropriate disclosures built-in, modeling potential adverse feedback early in the design process, followed by specific live testing and iteration for a small fraction of traffic before full deployment, and engaging with a diverse set of users and use-case scenarios.

2. Identify multiple metrics to assess training and monitoring: The use of several metrics rather than a single one will help to understand tradeoffs between different kinds of errors and experiences, e.g. considering metrics including feedback from user surveys, quantities that track overall system performance and short- and long-term product heath (e.g., click-through rate and customer lifetime value, respectively), and false positive and false negative rates sliced across different subgroups. Also, must be ensured that the metrics are appropriate for the context and goals of the system.

3. When possible, directly examine the raw data. ML models will reflect the data they are trained on, so analyzing the raw data carefully is essential to ensure you understand it. In cases where this is not possible, e.g., with sensitive raw data, understand the input data as much as possible while respecting privacy, e.g. by computing aggregate, anonymized summaries.

4. Understand the limitations of your dataset and model, e.g. a model trained to detect correlations should not be used to make causal inferences. Communicate limitations to users where possible.

5. Test: Learn from software engineering best test practices and quality engineering to make sure the AI system is working as intended and can be trusted. E.g. conduct rigorous unit tests to test each component of the system in isolation, conduct integration tests to understand how individual ML components interact with other parts of the overall system, and conduct iterative user testing to incorporate a diverse set of users' needs in the development cycles.

6. Continue to monitor and update the system after deployment: Continued monitoring will ensure the model takes real-world performance and user feedback (e.g., happiness tracking surveys, HEART framework) into account. Issues will occur, consider both short- and long-term solutions to issues. Before updating a deployed model, analyze how the candidate and deployed models differ, and how the update will affect the overall system quality and user experience.

3.6. Privacy, security and safety

AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented. Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must be ensured, considering the quality and integrity of the data, and ensuring legitimized access to data. (European Commission 2019.)

Privacy is especially critical for AI since the sophisticated insights generated by AI systems often stem from data that is detailed and personal. Trustworthy AI must comply with data regulations and only use data for the stated and agreed-upon purposes. Companies need to know what customer data is being collected and why, and whether the data is being used in the way customers understand and agree. *Customers should be given the required level of control over their data, including the ability to opt in or opt out of having their data shared.* If customers have concerns about data privacy, they need an avenue to voice those concerns (Saif & Ammanath 2020.) Recommended practices for privacy according to Google (Responsible AI practices 2019) are 1. collect and handle data responsibly, 2. leverage on-device processing where appropriate, and 3. appropriately safeguard the privacy of ML models.

AI must be protected from cybersecurity risks that might lead to physical and/or digital harm. To help ensure the safety and security of the AI systems, companies need to identify potential threats and address all kinds of risks—external, physical, and digital among many others—and then communicate those risks to users, in addition to develop and approach to combat these threats. Although external risks tend to get the most attention, internal risks such as fraud can be just as serious. For each AI use case, companies need to assess whether the potential benefits sufficiently outweigh the associated risks. (Saif & Ammanath 2020, Google Responsible AI practices 2019.)

3.7. Societal AI (AI for Social Good)

Hager et al. (2017) state that the term "social good" is intended to focus AI research on areas that are to benefit a broad population without direct economic impact or return. Research led from applications, from actual use, is important for this area of work, and in shaping AI for Social Good. Use inspired work in this area will lead to questions that are crucial to making a social impact. Research on AI for Social Good is *closely related to the basic principles of human-centered AI stressing consequences of a certain solutions*: many of these applications are seen to be decision aids, assisting the human. Research on AI for Social Good require research with interdisciplinary teams, where part of the team is rooted firmly in the domain discipline. A novel aspect of this interdisciplinary work is new methods for evaluating interdisciplinary work and measuring impact. Interpretability and transparency of the algorithms will remain key requirements in the future of AI for Social Good.

Crowley et al. 2019 define societal AI as an ability to model and understand the consequences of complex network effects in large-scale mixed communities of humans and AI systems interacting over various temporal and spatial scales. This

includes the ability to balance requirements related to individual users and the common good and societal concerns. In addition to human wellbeing and benefits, European Commission (Guidelines for Trustworthy AI 2019) underlines the *environmental impact of AI solutions*: they must be sustainable and environmentally friendly, and they should consider the environment, including other living beings.

McKinseys' report 'Applying AI for social good' (Chui et al. 2018) identifies 18 potential bottlenecks based on interviews with social-domain experts and AI researchers and practitioners. According to this report, the most significant bottlenecks in societal AI are data accessibility: data needed for social impact uses may not be easily accessible. Other identified challenges are a shortage of talent to develop AI solutions (not enough available AI expertise), and "last-mile" implementation challenges (training AI models is in short supply).

Societal AI typically refers to NGOs and other social-sector organizations, which brings certain challenges especially in the implementation of AI solutions. For example, hand-off might fail due to technical problems when deploying and sustaining AI models that require AI-related skills. In addition, organizations may have difficulties to interpret the results of an AI model. Even if a model achieves a desired level of accuracy on test data, new or unanticipated failure cases often appear in real-life scenarios. An understanding of how the solution works may require a data scientist or "translator." McKinseys' report (Chui et al. 2018) represents relevant risks similar to identified risks in human-centered AI including areas e.g. bias and fairness, privacy, safe use and security and explainability.

3.8. Trust

Trust is the willingness of a user to invest in an emotional bond with the system, and it is predicated on security of the system's data, the feeling of human control, and the quality of the results the system provides (IBM Design for AI 2019). Control is a backbone for trust (esp. in automated systems): when users have the right level of control over the system, they're more likely to trust it (unlike general definition of trust in technology, where user lack control and is thus vulnerable over the outcome of the trusted party). For AI to be trustworthy, all participants have a right to understand how their data is being used and how the AI is making decisions. In addition, trustworthy AI systems need to include *policies that clearly establish who is responsible and accountable for their output.* An organizational structure and policies should be put in place that can help clearly determine who is responsible for the output of AI system decisions. (Saif & Ammanath, 2020.) Thus, trust is the outcome of a human-centered AI design where e.g. issues in explainability, transparency and accountability are considered successfully.

Reliability or predictability (belief that the technology will consistently operate properly (see e.g. McKnight et al. 2011) is a critical factor in trust in technology. To be trustworthy, AI must scale up well and generate consistent and reliable outputs—performing tasks properly in less-than-ideal conditions and when encountering unexpected situations and data. If AI fails, it must fail in a predictable, expected manner. The human factor is a critical element: understanding how human input affect reliability, determining who are the right people to provide input, and ensuring those people are properly equipped and trained—particularly regarding bias and ethics. (Saif & Ammanath, 2020.)

Interestingly, Google states in their People + AI Guidebook (2019) that user shouldn't completely trust the system: *based on system explanations, the user should know when to trust the system's predictions and when to apply their own judgement.* This is an interesting statement considering both the human-in-the-loop philosophy and the previous research about the importance of trust in technology adoption process (see e.g. Gefen et al. 2003).

References

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. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019). Retrieved from https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

Burt, A. (2019). The AI Transparency Paradox. Retrieved from https://hbr.org/2019/12/the-ai-transparency-paradox

Chui et al. (2018). Applying artificial intelligence for social good. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Deloitte (2019). Transparency and Responsibility in Artificial Intelligence – a call for explainable AI. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf

Engler, A. (2020) The case for AI transparency requirements. Brookings Institution's Artificial Intelligence. Retrieved from https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/

European Commission (2019). Ethics guidelines for trustworthy AI. Shaping Europe's digital future. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Parliament (2019). Understanding algorithmic decision-making: Opportunities and challenges. EPRS | European Parliamentary Research Service. Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem, Information, Communication & Society, 20:1, 63-80. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/Gillespie_2017_Algorithmically-recognizable.pdf

Google (2019). People + AI Guidebook. Retrieved from https://pair.withgoogle.com/guidebook/

Google (2019). Responsible AI practices. Retrieved from https://ai.google/responsibilities/responsible-ai-practices/

Hager et al. (2017). Artificial Intelligence for Social Good. Computing Community Consortium (CCC). Retrieved from https://arxiv.org/pdf/1901.05406.pdf

Hulkko, V. (2018). Most value from AI with human-in-the-loop solutions
Retrieved from https://silo.ai/most-value-human-in-the-loop-ai/

Crowley et al. (2019). Toward AI Systems that Augment and Empower Humans by Understanding Us, our Society and the World Around Us. HumaneAI. Retrieved from https://www.humane-ai.eu/wp-content/uploads/2019/11/D21-HumaneAI-Concept.pdf

Hume, K. (2018). When Is It Important for an Algorithm to Explain Itself? Retrieved from https://hbr.org/2018/07/when-is-it-important-for-an-algorithm-to-explain-itself

IBM (2019). Design for AI. Retrieved from https://www.ibm.com/design/ai/

Kwan, N. (2018). The Hidden Dangers in Algorithmic Decision Making. Retrieved from https://towardsdatascience.com/the-hidden-dangers-in-algorithmic-decision-making-27722d716a49

Li, F.F., Etchemendy, J. A. (2018). Common goal for the brightest minds from Stanford and beyond: Putting humanity at the center of AI. Retrieved from https://hai.stanford.edu/news/introducing-stanfords-human-centered-ai-initiative

Lovejoy, J. (2019). Human-centered AI cheat-sheet. Retrieved from https://uxdesign.cc/human-centered-ai-cheat-sheet-1da130ba1bab

Martin, K. (2018). Ethical Implications and Accountability of Algorithms. J Bus Ethics 160, 835–850. Retrieved from https://link.springer.com/article/10.1007/s10551-018-3921-3#citeas

Matsakis, L. (2018). What Does a Fair Algorithm Actually Look Like? Retrieved from https://www.wired.com/story/what-does-a-fair-algorithm-look-like/

MindAI (2018) Lack of transparency could be AI's fatal flaw. Retrieved from https://medium.com/mind-ai/lack-of-transparency-could-be-ais-fatal-flaw-7c33b855928c

Mintz, S. (2020). Ethical AI is Built On Transparency, Accountability and Trust. Corporate Compliance Insights. Retrieved from https://www.corporatecomplianceinsights.com/ethical-use-artificial-intelligence/

Nushi, B. (2020). How to build effective human-AI interaction: Considerations for machine learning and software engineering. Retrieved from https://towardsdatascience.com/how-to-build-effective-human-ai-interaction-considerations-for-machine-learning-and-software-409838d9b358

Ramamoorthy, R. (2019). How Can We Remove the Bias in Algorithmic Decision Making? Retrieved from https://www.dataversity.net/how-can-we-remove-the-bias-in-algorithmic-decision-making/

Reese, B. (?) Explainable Artificial Intelligence v1.0. A Deep Dive Into XAI. Retrieved from: https://gigaom.com/deepdive/explainable-artificial-intelligence/

Rossi, F., Sekaran, A., Spohrer, J., Caruthers, R. (2019). IBM Everyday Ethics for Artificial Intelligence. Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Saif , I., Ammanath, B. (2020). 'Trustworthy AI' is a framework to help manage unique risk. Deloitte AI Institute. In MIT Technology Review March 25, 2020. Retrieved from https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

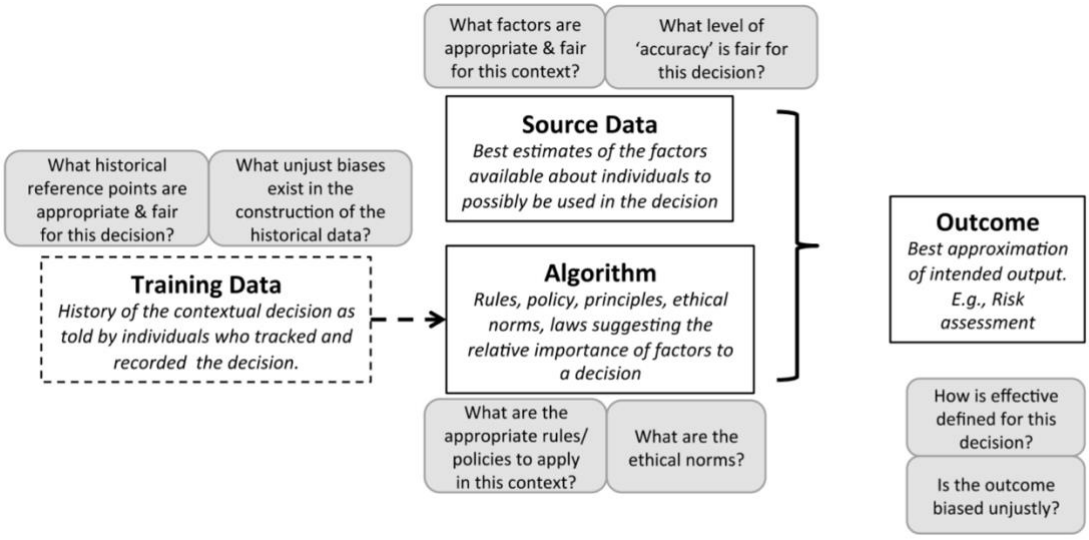
Sondergraad, P. (2019). AI Governance – what are the KPIs? Retrieved from https://2021.ai/ai-governance-kpi/

Wachter, S., Mittelstadt B., Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 2018. Retrieved from https://arxiv.org/abs/1711.00399

Whittaker et al. (2018). AI Now Report 2018. Retrieved from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Wärnestål, P. (2019). Human-Centered AI – the impact on design. Retrieved from https://www.inuse.se/blogg/human-centered-ai-impact-design/

Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interaction. Retrieved from https://interactions.acm.org/archive/view/july-august-2019/toward-human-centered-ai



Transparency model proposal (Martin 2018)

HCAI terminology