Human-Centered Artificial Intelligence Maturity Model for AI developers

Artificial intelligence has become increasingly prevalent in everyday life in ways that have positive and negative consequences for users, other people, and society. Reflecting the increasing importance and integration of AI in people's lives, there is a move towards human-centered AI (HCAI), which has the goal of placing the human rather than technology at the center of AI development.

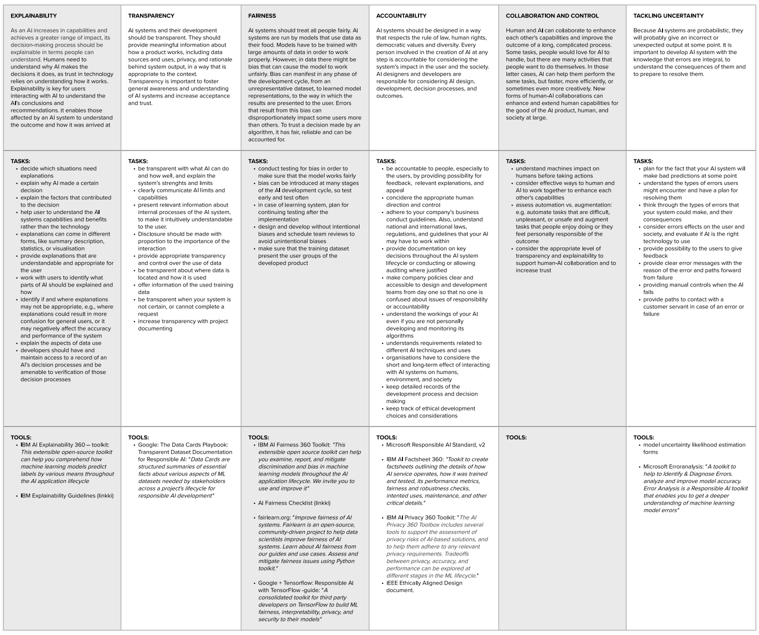
Human-centered AI aims to bring efficient solutions to user problems and provide positive and beneficial outcomes to the users, to those affected by their operation, and to society in general. It refers to development of AI systems that are trustworthy and ethical. In addition, HCAI seeks human-friendly collaboration in mixed human-AI settings preserving human control. HCAI is also about managing the unpredictability of AI. HCAI and its requirements and building blocks might be unfamiliar to ai developers, as they are still finding their ways to work with ai. the use of ai might bring new ai-related factors and requirements that should be acknowledged in the development. We are aiming to increase knowledge of HCAI requirements in AI developer companies with a maturity model that is appropriate to practical use, gives comprehensive guidance, and provides helpful tools and toolkits to promote the practical implementation of the model.

HCAI related dimension recognised from related literature:

- working with AI uncertainty
- user control and human-AI collaboration
- ethical development and use of AI: transparency, accountability, and fairness
- trustworthiness of AI: explainability & transparency to build trust between the user and AI

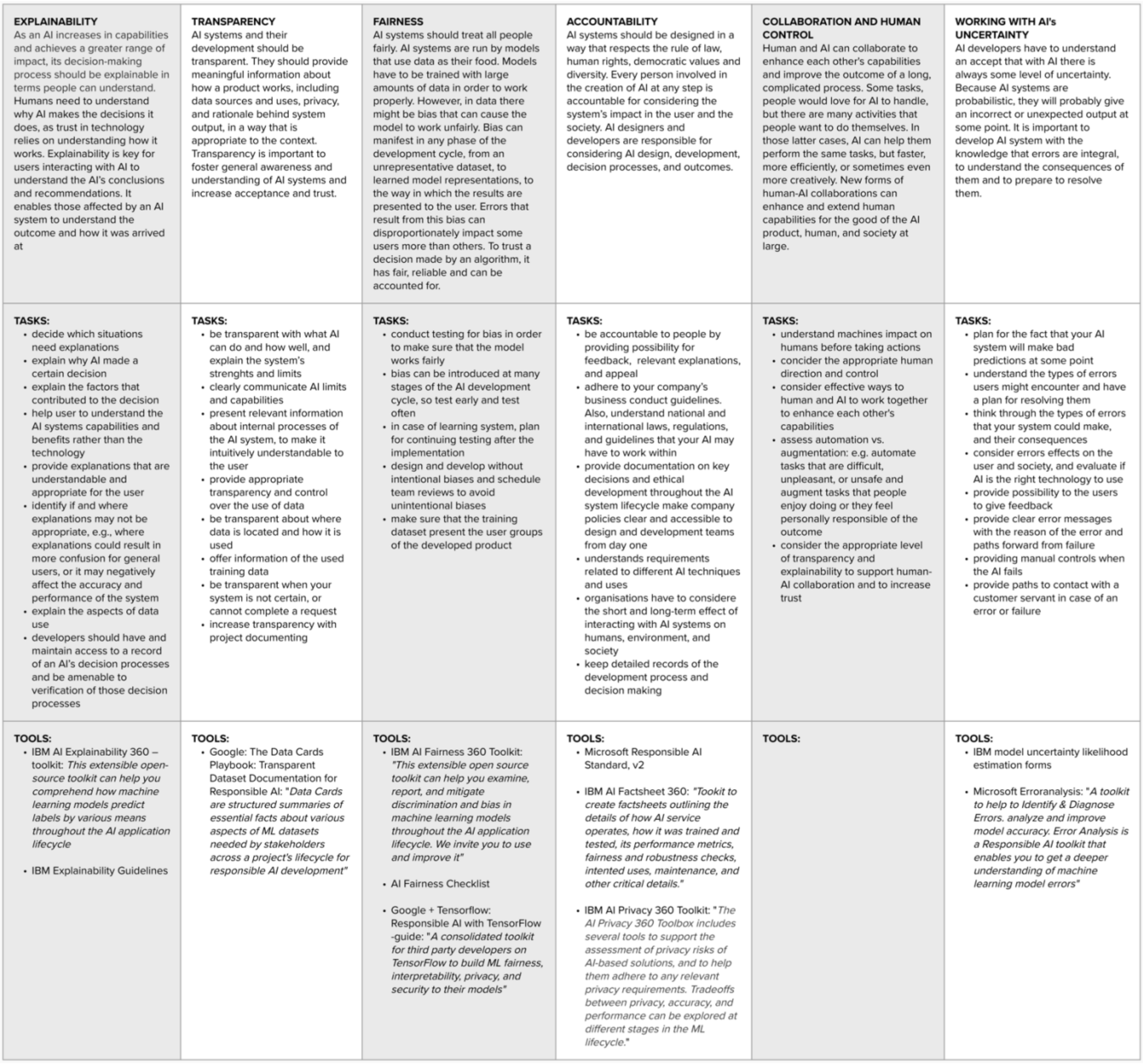
Model structure:

- short introduction on the HCAI dimension
- each dimension is further specified by tasks – practices or activities related to the HCAI dimension
- tools to support the development



EXPLAINABILITY	TRANSPARENCY	FAIRNESS	ACCOUNTABILITY	COLLABORATION AND CONTROL	TACKLING UNCERTAINTY
As an AI increases in capabilities and achieves a greater range of impact, its decision-making process should be explainable in terms people can understand. Humans need to understand why AI makes the decisions it does, as trust in technology relies on understanding how it works. Explainability is key for users interacting with AI to understand the AI's conclusions and recommendations. It enables those affected by an AI system to understand the outcome and how it was arrived at	AI systems and their development should be transparent. They should provide meaningful information about how a product works, including data sources and uses, privacy, and rationale behind system output, in a way that is appropriate to the context. Transparency is important to foster general awareness and understanding of AI systems and increase acceptance and trust.	AI systems should treat all people fairly. AI systems are run by models that use data as their food. Models have to be trained with large amounts of data in order to work properly. However, in data there might be bias that can cause the model to work unfairly. Bias can manifest in any phase of the development cycle, from an unrepresentative dataset, to learned model representations, to the way in which the results are presented to the user. Errors that result from this bias can disproportionately impact some users more than others. To trust a decision made by an algorithm, it has fair, reliable and can be accounted for.	AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity. Every person involved in the creation of AI at any step is accountable for considering the system's impact in the user and the society. AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes.	Human and AI can collaborate to enhance each other's capabilities and improve the outcome of a long, complicated process. Some tasks, people would love for AI to handle, but there are many activities that people want to do themselves. In those latter cases, AI can help them perform the same tasks, but faster, more efficiently, or sometimes even more creatively. New forms of human-AI collaborations can enhance and extend human capabilities for the good of the AI product, human, and society at large.	Because AI systems are probabilistic, they will probably give an incorrect or unexpected output at some point. It is important to develop AI system with the knowledge that errors are integral, to understand the consequences of them and to prepare to resolve them.
TASKS: • decide which situations need explanations • explain why AI made a certain decision • explain the factors that contributed to the decision • help user to understand the AI systems capabilities and benefits rather than the technology • explanations can come in different forms, like summary description, statistics, or visualisation • provide explanations that are understandable and appropriate for the user • work with users to identify what parts of AI should be explained and how • identify if and where explanations may not be appropriate, e.g., where explanations could result in more confusion for general users, or it may negatively affect the accuracy and performance of the system • explain the aspects of data use • developers should have and maintain access to a record of an AI's decision processes and be amenable to verification of those decision processes	**TASKS:** • be transparent with what AI can do and how well, and explain the system's strenghts and limits • clearly communicate AI limits and capabilities • present relevant information about internal processes of the AI system, to make it intuitively understandable to the user. • Disclosure should be made with proportion to the importance of the interaction • provide appropriate transparency and control over the use of data • be transparent about where data is located and how it is used • offer information of the used training data • be transparent when your system is not certain, or cannot complete a request • increase transparency with project documenting	**TASKS:** • conduct testing for bias in order to make sure that the model works fairly • bias can be introduced at many stages of the AI development cycle, so test early and test often • in case of learning system, plan for continuing testing after the implementation • design and develop without intentional biases and schedule team reviews to avoid unintentional biases • make sure that the training dataset present the user groups of the developed product	**TASKS:** • be accountable to people, especially to the users, by providing possibility for feedback, relevant explanations, and appeal • concidere the appropriate human direction and control • adhere to your company's business conduct guidelines. Also, understand national and international laws, regulations, and guidelines that your AI may have to work within • provide documentation on key decisions throughout the AI system lifecycle or conducting or allowing auditing where justified • make company policies clear and accessible to design and development teams from day one so that no one is confused about issues of responsibility or accountability • understand the workings of your AI even if you are not personally developing and monitoring its algorithms • understands requirements related to different AI techniques and uses • organisations have to considere the short and long-term effect of interacting with AI systems on humans, environment, and society • keep detailed records of the development process and decision making • keep track of ethical development choices and considerations	**TASKS:** • understand machines impact on humans before taking actions • consider effective ways to human and AI to work together to enhance each other's capabilities • assess automation vs. augmentation: e.g. automate tasks that are difficult, unpleasant, or unsafe and augment tasks that people enjoy doing or they feel personally responsible of the outcome • consider the appropriate level of transparency and explainability to support human-AI collaboration and to increase trust	**TASKS:** • plan for the fact that your AI system will make bad predictions at some point • understand the types of errors users might encounter and have a plan for resolving them • think through the types of errors that your system could make, and their consequences • consider errors effects on the user and society, and evaluate if AI is the right technology to use • provide possibility to the users to give feedback • provide clear error messages with the reason of the error and paths forward from failure • providing manual controls when the AI fails • provide paths to contact with a customer servant in case of an error or failure
TOOLS: • IBM AI Explainability 360 – toolkit: *This extensible open-source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle* • IBM Explainability Guidelines (linkki)	**TOOLS:** • Google: The Data Cards Playbook: Transparent Dataset Documentation for Responsible AI: "*Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a project's lifecycle for responsible AI development*"	**TOOLS:** • IBM AI Fairness 360 Toolkit: *"This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it"* • AI Fairness Checklist (linkki) • fairlearn.org: "*Improve fairness of AI systems. Fairlearn is an open-source, community-driven project to help data scientists improve fairness of AI systems. Learn about AI fairness from our guides and use cases. Assess and mitigate fairness issues using Python toolkit."* • Google + Tensorflow: Responsible AI with TensorFlow -guide: "*A consolidated toolkit for third party developers on TensorFlow to build ML fairness, interpretability, privacy, and security to their models"*	**TOOLS:** • Microsoft Responsible AI Standard, v2 • IBM AI Factsheet 360: "*Tookit to create factsheets outlining the details of how AI service operates, how it was trained and tested, its performance metrics, fairness and robustness checks, intented uses, maintenance, and other critical details."* • IBM AI Privacy 360 Toolkit: "*The AI Privacy 360 Toolbox includes several tools to support the assessment of privacy risks of AI-based solutions, and to help them adhere to any relevant privacy requirements. Tradeoffs between privacy, accuracy, and performance can be explored at different stages in the ML lifecycle.*" • IEEE Ethically Aligned Design document.	**TOOLS:**	**TOOLS:** • model uncertainty likelihood estimation forms • Microsoft Erroranalysis: "*A toolkit to help to Identify & Diagnose Errors. analyze and improve model accuracy. Error Analysis is a Responsible AI toolkit that enables you to get a deeper understanding of machine learning model errors*"

Human-Centered Artificial Intelligence Maturity Model for AI developers



EXPLAINABILITY As an AI increases in capabilities and achieves a greater range of impact, its decision-making process should be explainable in terms people can understand. Humans need to understand why AI makes the decisions it does, as trust in technology relies on understanding how it works. Explainability is key for users interacting with AI to understand the AI's conclusions and recommendations. It enables those affected by an AI system to understand the outcome and how it was arrived at	**TRANSPARENCY** AI systems and their development should be transparent. They should provide meaningful information about how a product works, including data sources and uses, privacy, and rationale behind system output, in a way that is appropriate to the context. Transparency is important to foster general awareness and understanding of AI systems and increase acceptance and trust.	**FAIRNESS** AI systems should treat all people fairly. AI systems are run by models that use data as their food. Models have to be trained with large amounts of data in order to work properly. However, in data there might be bias that can cause the model to work unfairly. Bias can manifest in any phase of the development cycle, from an unrepresentative dataset, to learned model representations, to the way in which the results are presented to the user. Errors that result from this bias can disproportionately impact some users more than others. To trust a decision made by an algorithm, it has fair, reliable and can be accounted for.	**ACCOUNTABILITY** AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity. Every person involved in the creation of AI at any step is accountable for considering the system's impact in the user and the society. AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes.	**COLLABORATION AND HUMAN CONTROL** Human and AI can collaborate to enhance each other's capabilities and improve the outcome of a long, complicated process. Some tasks, people would love for AI to handle, but there are many activities that people want to do themselves. In those latter cases, AI can help them perform the same tasks, but faster, more efficiently, or sometimes even more creatively. New forms of human-AI collaborations can enhance and extend human capabilities for the good of the AI product, human, and society at large.	**WORKING WITH AI's UNCERTAINTY** AI developers have to understand an accept that with AI there is always some level of uncertainty. Because AI systems are probabilistic, they will probably give an incorrect or unexpected output at some point. It is important to develop AI system with the knowledge that errors are integral, to understand the consequences of them and to prepare to resolve them.
TASKS: • decide which situations need explanations • explain why AI made a certain decision • explain the factors that contributed to the decision • help user to understand the AI systems capabilities and benefits rather than the technology • provide explanations that are understandable and appropriate for the user • identify if and where explanations may not be appropriate, e.g., where explanations could result in more confusion for general users, or it may negatively affect the accuracy and performance of the system • explain the aspects of data use • developers should have and maintain access to a record of an AI's decision processes and be amenable to verification of those decision processes	**TASKS:** • be transparent with what AI can do and how well, and explain the system's strenghts and limits • clearly communicate AI limits and capabilities • present relevant information about internal processes of the AI system, to make it intuitively understandable to the user • provide appropriate transparency and control over the use of data • be transparent about where data is located and how it is used • offer information of the used training data • be transparent when your system is not certain, or cannot complete a request • increase transparency with project documenting	**TASKS:** • conduct testing for bias in order to make sure that the model works fairly • bias can be introduced at many stages of the AI development cycle, so test early and test often • in case of learning system, plan for continuing testing after the implementation • design and develop without intentional biases and schedule team reviews to avoid unintentional biases • make sure that the training dataset present the user groups of the developed product	**TASKS:** • be accountable to people by providing possibility for feedback, relevant explanations, and appeal • adhere to your company's business conduct guidelines. Also, understand national and international laws, regulations, and guidelines that your AI may have to work within • provide documentation on key decisions and ethical development throughout the AI system lifecycle make company policies clear and accessible to design and development teams from day one • understands requirements related to different AI techniques and uses • organisations have to considere the short and long-term effect of interacting with AI systems on humans, environment, and society • keep detailed records of the development process and decision making	**TASKS:** • understand machines impact on humans before taking actions • concider the appropriate human direction and control • consider effective ways to human and AI to work together to enhance each other's capabilities • assess automation vs. augmentation: e.g. automate tasks that are difficult, unpleasant, or unsafe and augment tasks that people enjoy doing or they feel personally responsible of the outcome • consider the appropriate level of transparency and explainability to support human-AI collaboration and to increase trust	**TASKS:** • plan for the fact that your AI system will make bad predictions at some point • understand the types of errors users might encounter and have a plan for resolving them • think through the types of errors that your system could make, and their consequences • consider errors effects on the user and society, and evaluate if AI is the right technology to use • provide possibility to the users to give feedback • provide clear error messages with the reason of the error and paths forward from failure • providing manual controls when the AI fails • provide paths to contact with a customer servant in case of an error or failure
TOOLS: • IBM AI Explainability 360 – toolkit: *This extensible open-source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle* • IBM Explainability Guidelines	**TOOLS:** • Google: The Data Cards Playbook: Transparent Dataset Documentation for Responsible AI: "*Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a project's lifecycle for responsible AI development"*	**TOOLS:** • IBM AI Fairness 360 Toolkit: *"This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it"* • AI Fairness Checklist • Google + Tensorflow: Responsible AI with TensorFlow -guide: "*A consolidated toolkit for third party developers on TensorFlow to build ML fairness, interpretability, privacy, and security to their models"*	**TOOLS:** • Microsoft Responsible AI Standard, v2 • IBM AI Factsheet 360: *"Tookit to create factsheets outlining the details of how AI service operates, how it was trained and tested, its performance metrics, fairness and robustness checks, intented uses, maintenance, and other critical details."* • IBM AI Privacy 360 Toolkit: "*The AI Privacy 360 Toolbox includes several tools to support the assessment of privacy risks of AI-based solutions, and to help them adhere to any relevant privacy requirements. Tradeoffs between privacy, accuracy, and performance can be explored at different stages in the ML lifecycle."*	**TOOLS:**	**TOOLS:** • IBM model uncertainty likelihood estimation forms • Microsoft Erroranalysis: "*A toolkit to help to Identify & Diagnose Errors. analyze and improve model accuracy. Error Analysis is a Responsible AI toolkit that enables you to get a deeper understanding of machine learning model errors"*