Everyday Ethics for Artificial Intelligence

Acknowledgements

Francesca Rossi
AI Ethics Global Leader,
IBM Fellow
IBM Research

Noah Treviño
IBM Design
Program Office
Visual Designer

Almas Ahmed
IBM Blue Studio
Designer

Everyday Ethics for Artificial Intelligence

IBM Design Program Office

If you have questions, comments or suggestions please email edethics@us.ibm.com to contribute to this effort.

Adam Cutler
IBM Distinguished Designer,
IBM Design for AI

Milena Pribić
Senior Designer,
IBM Design for AI

Table of Contents

Ethics must be embedded in the design and development process from the very beginning of AI creation. Everyday Ethics for AI is meant to guide team discussions and daily practices.

"You can use an eraser on the drafting table or a sledgehammer on the construction site."

– Frank Lloyd Wright

IBM embraces five foundational pillars of trustworthy AI: Explainability, Fairness, Robustness, Transparency, and Privacy.[1] These pillars underpin the development, deployment and use of AI systems. This document and IBM's trustworthy AI pillars are meant to help you align on both process and outcomes.

This is a living document. Please experiment, play, use, and break what you find here and send us your feedback.

Designers and developers of AI systems are encouraged to be aware of these concepts and seize opportunities to put these ideas into practice. As you work with your team and others, please share this guide with them.

Introduction

Ethical decision-making isn't just another form of technical problem solving.

As AI designers and developers, we hold a vast share of the collective influence. We are creating systems that will impact millions of people. AI is rapidly growing in capability, impact and influence. As designers and developers of AI systems, it is an imperative to understand the ethical considerations of our work.

A technology-centric focus that solely revolves around improving the capabilities of an intelligent system doesn't sufficiently consider human needs. A trustworthy, human-centric AI system must be designed and developed in a manner that is aligned with the values and principles of the society or community it affects.

Ethics is a set of moral principles which help us discern between right and wrong. AI ethics is a set of guidelines that advise on the design, development, and use of artificial intelligence.[2] To create and foster trust between humans and machines, you must understand the ethical resources and standards available for reference during the design, building, and maintenance of AI. The large-scale focus on AI ethics by groups like the World Economic Forum, the Future of Privacy Forum, the Partnership on AI, and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, should be mirrored in businesses and working groups of all sizes.

The criteria and metrics for trustworthy AI systems will ultimately depend on the industry and use case they operate within. We hope this document serves as a central source that helps teams establish best practices. Designers and developers should never work in a vacuum and must stay in tune with users' needs and concerns. Constant improvement and assessment is key to ensuring that design and development teams address users' concerns. This document provides teams with a starting point and will surely evolve as AI capabilities continue to grow.

This guide provides discussion points concerning:

- Specific virtues and practices to promote.
- Guidance for designers and developers building and training AI.

Five Practices of Everyday Ethics

It's our collective responsibility to understand and evolve these ethical practices as AI capabilities increase over time. These practices provide an intentional framework for building and using AI systems alongside IBM's five pillars of trustworthy AI.

Take accountability for the outcomes of your AI system in the real world, no matter your role.

Be sensitive to a wide range of cultural norms and values, not just your own.

Work with your team to identify and address biases and promote inclusive representation.

Ensure humans can perceive, detect, and understand an AI decision process.

Preserve and fortify users' power over their own data and its uses.

Designers and developers of AI who want to go deeper into these practices should consider IBM's *Team Essentials for AI* course that trains practitioners on relevant design thinking methods.[3]

In addition, IBM Research has made their Trustworthy AI toolkits publicly available for use by developers and data scientists.[4] They include:

01 AI Explainability 360:
This open source toolkit includes an extensive set of techniques as well as guidance on how to choose the explainability algorithm that's right for your use case.

02 AI Fairness 360:
This is an open source software toolkit that enables developers to use state-of-the-art algorithms to regularly check for unwanted biases from entering their machine learning pipeline and to mitigate any biases that are discovered.

03 AI FactSheets 360:
Similar to nutrition labels for food or information sheets for appliances, this project increases transparency so that AI consumers better understand how the AI model or service was created.

04 Adversarial Robustness Toolbox:
These tools enable developers and researchers to evaluate and defend machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

05 AI Privacy 360:
This 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.

Running Example

A hotel chain wants to embed AI into an in-room virtual assistant/concierge to augment and personalize their users' stays. We'll use the project team in charge of this effort as an example throughout the document. This conversational agent will include capabilities such as:

- Agentive-style assistance.
- Introduction to their room and services in their preferred language.
- Control of room facilities through natural language.
- Sending a request directly to the service team through the in-room virtual assistant.

Consider outcomes

Take accountability for the outcomes of your AI system in the real world, no matter your role.

Human judgment plays a role throughout a seemingly objective system of logical decisions. It is humans who write algorithms, who define success or failure, who make decisions about the uses of systems and who may be affected by a system's outcomes. Our level of accountability is also related to AI transparency, meaning everyone involved should be aware of what data is collected, how it's used and stored, and who has access to it. Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world, as are the companies invested in its development.

"AI recommendations give data points for managers to consider, but the decision-making and accountability remains with people."

– Responsible Use of Technology: The IBM Case Study, WEF[5]

Recommended actions to take

01
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. As an AI designer or developer, it is your responsibility to know.

02
Understand where the responsibility of the company/software ends. You may not have control over how data or a tool will be used by a user, client, or other external source.

03
Keep detailed records of your design processes and decision making. Determine a strategy for keeping records during the design and development process to encourage best practices and iteration.

04
Adhere to your company's business conduct guidelines. Also, understand national and international laws, regulations, and guidelines[6] that your AI system may have to work within. You can find other related resources in the IEEE Ethically Aligned Design document.[7]

To consider
Understand the workings of your AI system even if you're not personally developing and monitoring its algorithms.

The entire team should work together to choose robust components that minimize risks and enable users' confidence in system outcomes.[8]

Refer to secondary research by sociologists, linguists, behaviorists, and other professionals to understand ethical issues in a holistic context.

Questions for your team
How does accountability change according to the levels of user influence over an AI system?

Is the AI to be embedded in a human decision-making process, is it making decisions on its own, or is it a hybrid?

How will our team keep records of our process?

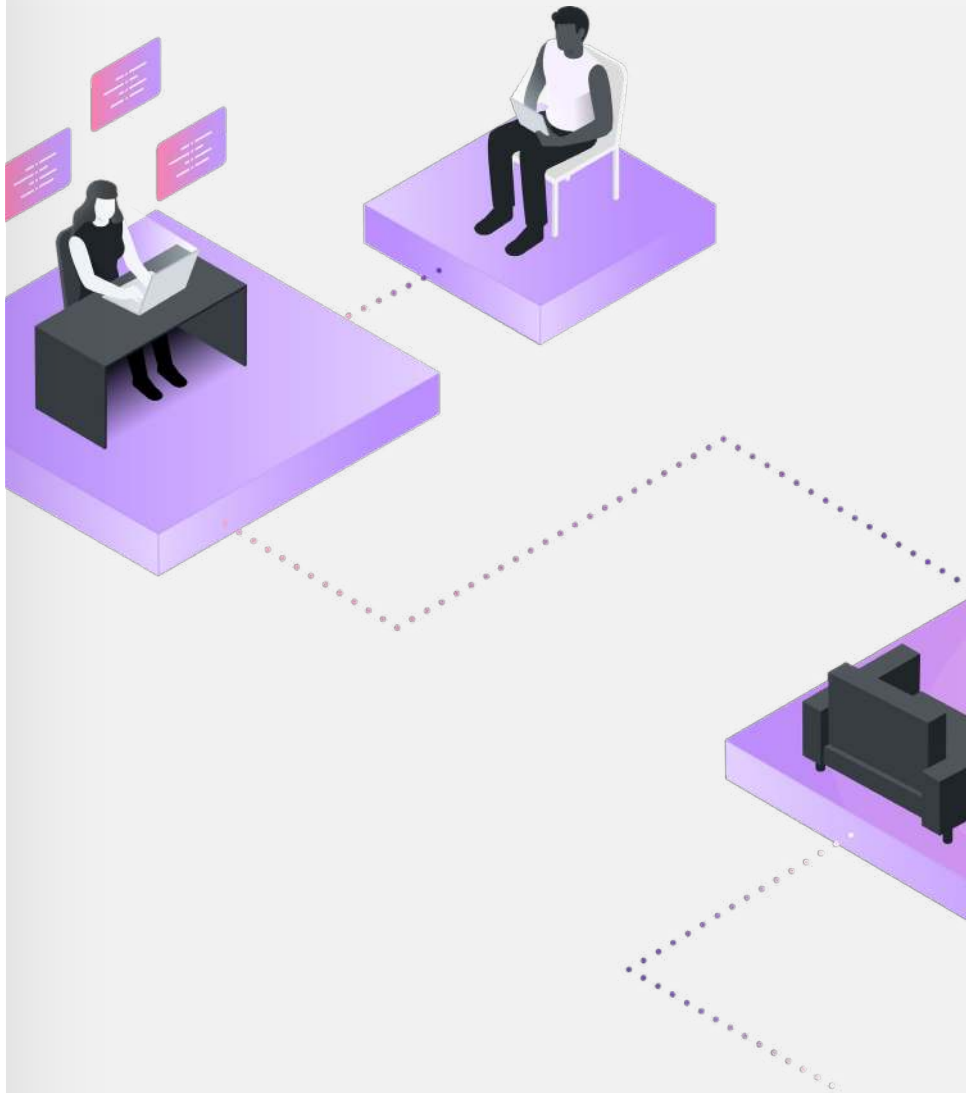
How do we keep track of ethical design choices and considerations after the launch of the AI system?

Will others new to our effort be able to understand our records?

Running Example

– The team utilizes design researchers to contact real guests in the hotels to understand their wants and needs through face-to-face user interviews.

– The team considers their own responsibility when the hotel assistant's feedback does not meet the needs or expectations of guests. They have implemented a feedback learning loop to better understand preferences and have highlighted the ability for a guest to turn off the AI assistant at any point during their stay.



Align with norms and values

Be sensitive to a wide range of cultural norms and values, not just your own.

AI works alongside diverse, human interests. People make decisions based on any number of contextual factors, including their experiences, memories, upbringing, and cultural norms. These factors allow us to have a fundamental understanding of "right and wrong" in a wide range of contexts, at home, in the office, or elsewhere. This is second nature for humans, as we have a wealth of experiences to draw upon.

Today's AI systems do not have these types of experiences to draw upon, so it is the responsibility of designers and developers to collaborate with each other in order to ensure consideration of existing values. Care is required to ensure sensitivity to a wide range of cultural norms and values. Companies advancing AI have an obligation to address these issues proactively.

As daunting as it may seem to take value systems into account, the common core of universal principles is that they are a cooperative phenomenon. Successful teams already understand that cooperation and collaboration leads to the best outcomes.

"With trust as the cornerstone of our leadership in AI innovation, IBM is the partner that businesses need right now as they look to use AI as a force for positive change. And at this moment in the long arc of human progress, that matters not just for our company, but for our customers and society at large."

– Principles and Practices for Building More Trustworthy AI[9]

Recommended actions to take

01
Consider the culture that establishes the value systems you're designing within. Whenever possible, bring in policymakers and academics that can help your team articulate relevant perspectives.

02
Work with design researchers to understand and reflect your users' values. You can find out more about this process on IBM's Design Research site.[10]

03
Consider mapping out your understanding of your users' values and aligning the AI system's actions accordingly with an Ethics Canvas.[11] Values will be specific to certain use cases and affected communities. Alignment will allow users to better understand your AI system's actions and intents.

To consider
If you need somewhere to start, consider IBM's five pillars of trustworthy AI, Principles of Trust and Transparency, Standards of Corporate Responsibility[12] or use your company's standards documentation.

Values are subjective and differ globally. Global companies must take into account linguistic barriers and cultural differences.

Well-meaning values can create unintended consequences. e.g. a tailored political newsfeed provides users with news that aligns with their beliefs but does not holistically represent the gestalt.

Questions for your team
Which group's values are expressed by our AI system and why?

How do we agree on which values to consider as a team?

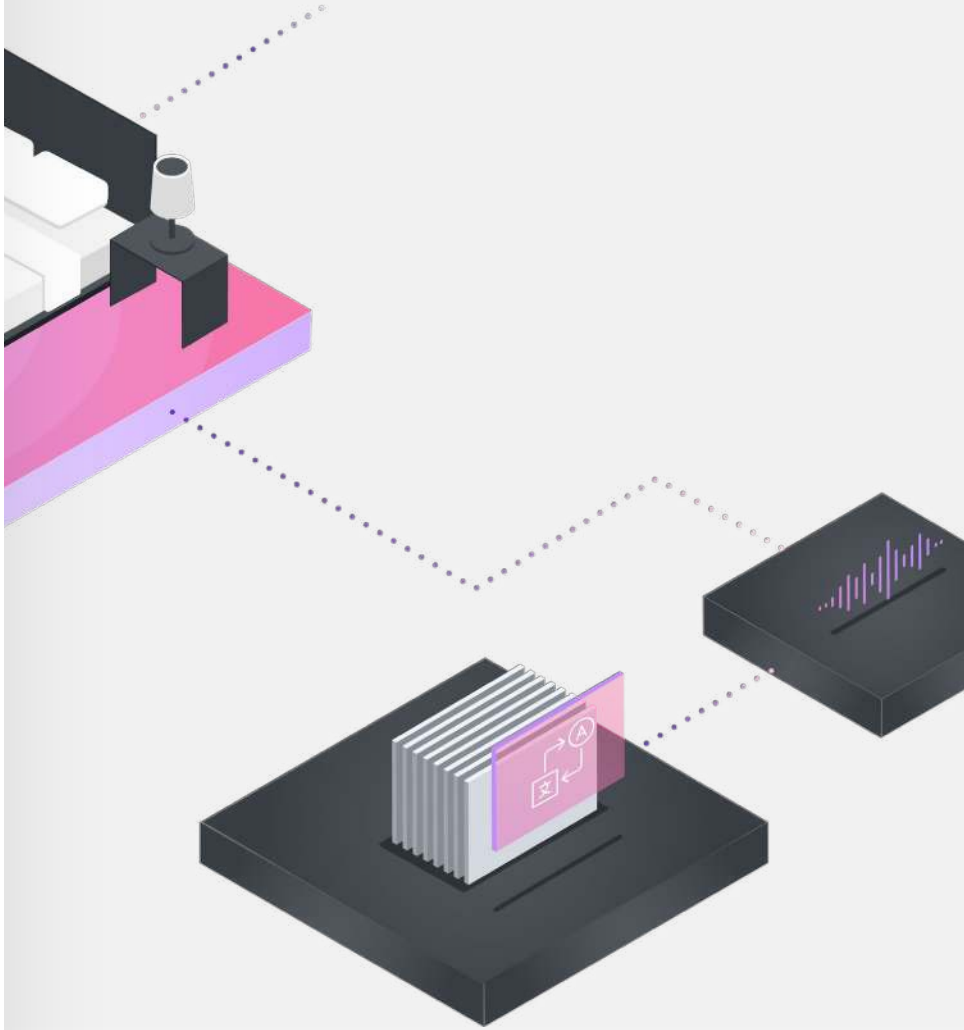
How do we change or adjust the values reflected by our AI system as our values evolve over time?

Running Example

- The team understands that for a voice-activated assistant to work properly, it must be “always listening” for a wake word. The team makes it clear to guests that the AI assistant is designed to not keep any data, or monitor guests, in both cases without their knowledge, even if it is listening for a wake word.

- The audio collected while listening for a wake word is auto-deleted every 5 seconds. Even if a guest opts in, the AI assistant does not actively listen in on guests unless it is called upon.

- The team knows that this agent will be used in hotels across the world, which will require different languages and customs. They consult with linguists to ensure the AI assistant will be able to speak in guests’ respective languages and respect applicable customs.



Minimize bias and improve inclusivity

Work with your team to identify and address biases and promote inclusive representation.

AI provides deeper insight into our personal lives when interacting with our sensitive data. As humans are inherently vulnerable to biases, and are responsible for building AI, there are chances for human bias to be embedded in the systems we create.

Bias can be present both in the algorithm of the AI system and in the data used to train and test it. It can emerge as a result of cultural, social, or institutional expectations.

Although bias can never be fully eliminated, it is the role of a responsible team to minimize algorithmic bias through ongoing research and responsible data collection representative of a diverse population.

"AI systems do not operate in isolation. They help people make decisions that directly affect other people's lives. If we are to develop trustworthy AI systems, we need to consider all the factors that can chip away at the public's trust in AI."

– Reva Schwartz, NIST[13]

Recommended actions to take

01
Real-time analysis of AI brings to light both intentional and unintentional biases. When bias in data becomes apparent, the team must investigate and understand where it originated and how it can be mitigated.

02
Design and develop without intentional biases and schedule team reviews to avoid unintentional biases. Unintentional biases can include stereotyping, confirmation bias, and sunk cost bias (see pages 34 and 35).

03
Utilize a feedback mechanism or open dialogue with users to raise awareness of user-identified biases or issues.

To consider
Diverse teams help to represent a wider variation of experiences to minimize bias. Embrace team members of different ages, ethnicities, genders, educational disciplines, and cultural perspectives.

Your AI system may be susceptible to different types of bias based on the type of data it ingests. Monitor training and results in order to quickly respond to issues. Test early and often for model robustness and performance.

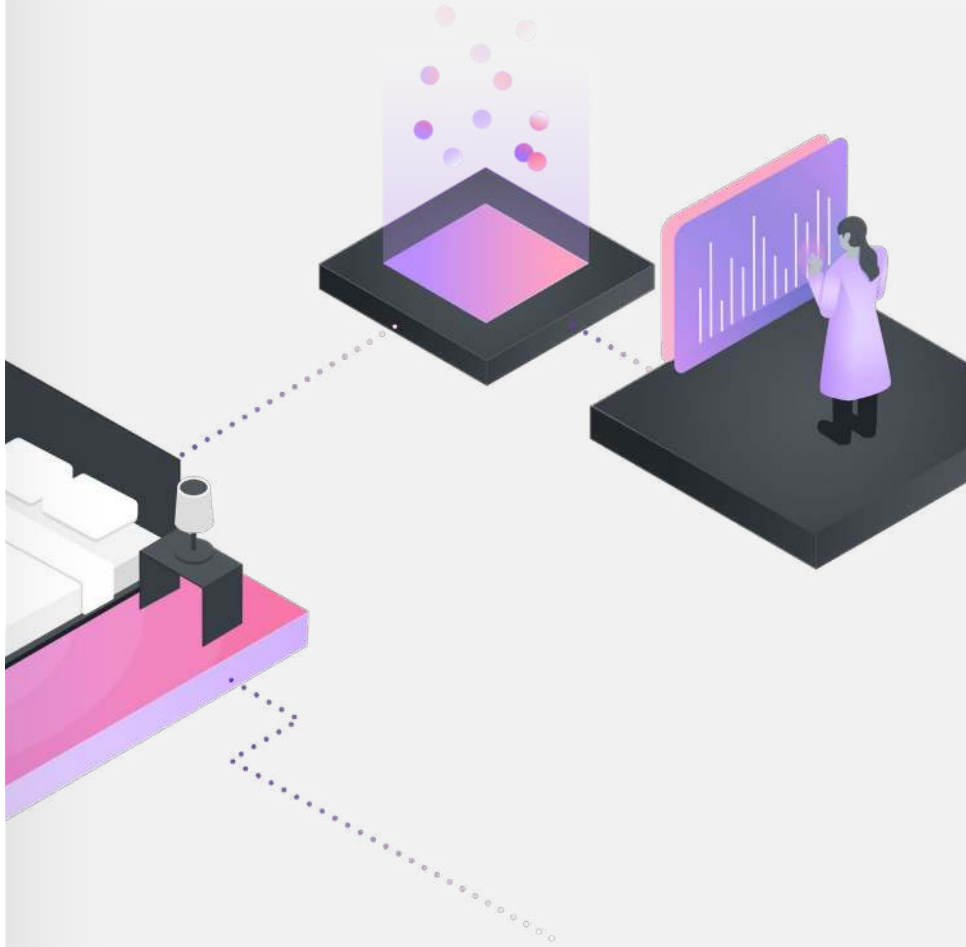
Questions for your team
How can we identify and audit unintentional biases that we run into during the design and development of our AI system?

The status quo changes over time. How do we instill methods to reflect that change in our ongoing data collection?

How do we best collect feedback from users in order to correct unintentional bias in design or decision-making?

Running Example

- The team ensures members of the hotel's global management that the data collected about a user's race, gender, etc. in combination with their usage of AI, will not be used to market to or exclude certain demographics.
- The team inherited a set of data about guests from the hotel. After analyzing this data and implementing it into a build of the agent, they realize that it has a degree of algorithmic bias from the data. The team starts over and retrains the model on a bigger, more diverse set of data.



Unconscious Bias Definitions

The average knowledge worker is unaware of the many different types of biases. While this list is not all-encompassing, these biases are some of the more common types to be consciously aware of when designing and developing for AI.

Shortcut Biases

"I don't have the time or energy to think about this."

Availability Bias
Overestimating events with greater "availability" in memory—influenced by how recent, unusual, or emotionally charged the memories may be.

Base Rate Fallacy
The tendency to ignore general information and focus on specific information (a certain case).

Congruence Bias
The tendency to test hypotheses exclusively through direct testing, instead of testing alternative hypotheses.

Empathy Gap Bias
The tendency to underestimate the influence or strength of feelings, in either ones' self or others.

Stereotyping
Expecting a member of a group to have certain characteristics without having actual information about that individual.

Impartiality Biases

"I know I'm wrong sometimes, but I'm right about this."

Anchoring Bias
To rely too much on one trait or piece of information when making decisions (usually the first piece of information that we acquire on that subject).

Bandwagon Bias
The tendency to do or believe things because many other people do (groupthink).

Bias Blind Spot
The tendency to see oneself as less biased than others, or to be able to identify more cognitive biases in others than in oneself.

Confirmation Bias
The tendency to search for, interpret, or focus on information in a way that confirms one's preconceptions.

Halo Effect
The tendency of an overall impression to influence the observer. Positive feelings in one area causes ambiguous or neutral traits to be viewed positively.

Self-Interest Biases

"We contributed the most. They weren't very cooperative."

Ingroup / Outgroup Bias
The tendency or pattern of favoring members of one's ingroup over outgroup members.

Sunk Cost Bias
The tendency to justify past choices, even though they no longer seem valid.

Status Quo Bias
The tendency to maintain the current situation — even when better alternatives exist.

Not Invented Here Bias
Aversion to contact with or use of products, research, standards, or knowledge developed outside a group.

Self-Serving Bias
The tendency to focus on strengths/achievements and overlook faults/failures. To take more responsibility for their group's work that they give to other groups.

Ensure explainability

Ensure humans can perceive, detect, and understand an AI decision process.

While transparency communicates the purpose and characteristics of an AI system, simple and straightforward explanations show users how and why a system arrived at a particular conclusion.

In general, we don't blindly trust those who can't explain their reasoning. The same goes for AI, perhaps even more so.[14] As an AI system increases in capabilities and achieves a greater range of impact, its decision-making process should be explainable in terms people can understand. Explainability is key for users interacting with AI to understand its conclusions and recommendations.

Your users should always be aware that they are interacting with AI. Good design does not sacrifice transparency and explainability while creating a seamless experience. Imperceptible AI is not ethical AI.

"Any AI system on the market that is making determinations or recommendations with potentially significant implications for individuals should be able to explain and contextualize how and why it arrived at a particular conclusion."

– IBM's Explainability Pillar[15]

Recommended actions to take

01
Allow for questions. A user should be able to ask why an AI system is doing what it's doing on an ongoing basis. This should be clear and up front in the user interface at all times.

02
Decision-making processes must be reviewable, especially if the AI system is working with highly sensitive personal information data like personally identifiable information, protected health information, and/or biometric data.

03
When an AI system is assisting users with making any highly sensitive decisions, it must be able to provide them with a sufficient explanation of recommendations, the data used, and the reasoning behind the recommendations.

04
Teams should have and maintain access to a record of an AI system's decision processes and be amenable to verification of those decision processes.

To consider
Explainability is needed to build public confidence in disruptive technology, to promote safer practices, and to facilitate broader societal adoption.

There are situations where users may not have access to the full decision process that an AI system might go through, e.g., financial investment algorithms.

Ensure an AI system's level of transparency is clear. Users should stay generally informed on the AI's intent even when they can't access a breakdown of the AI's process.

Questions for your team
How do we build explainability into our experience without detracting from user experience or distracting from the task at hand?

Do certain processes or pieces of information need to be hidden from users for security or IP reasons? How is this explained to users?

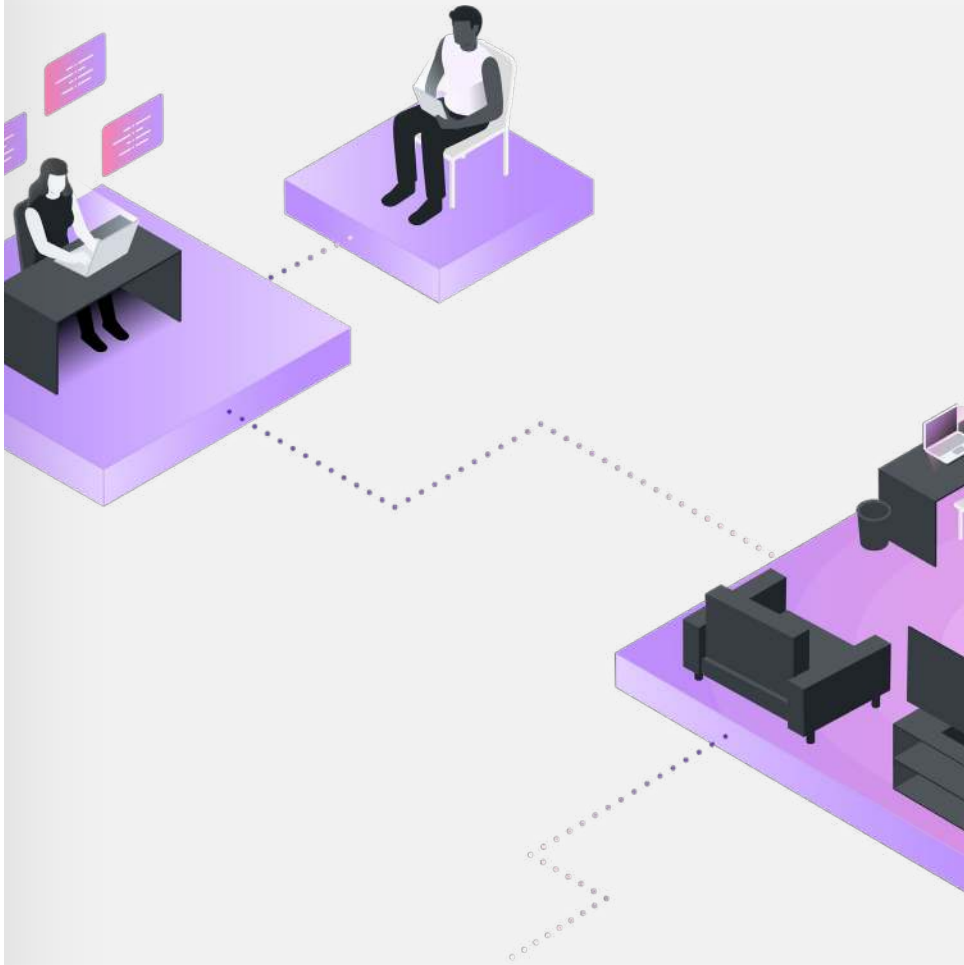
Which segments of the AI system's decision processes can be articulated for users in an easily digestible and explainable fashion?

Running Example

- Per GDPR (General Data Protection Regulation)[16], a guest must explicitly opt in to use the hotel room assistant. Additionally, they will be provided with a transparent UI in their room to show how the voice AI makes its recommendations and suggestions.

- A researcher on the team, through interviews with hotel guests, learns that the guests want a way to opt into having their personal information stored. The team provides consent mechanisms so that guests (through voice or graphic UI) can allow the system to store pieces of information.

- With permission, the AI assistant offers recommendations for places to visit during their stay. Guests can ask why these recommendations are made and which set of data is being utilized to make them.



Protect user data

Preserve and fortify users' power over their own data and its uses.

It's your team's responsibility to keep users empowered with control over their interactions and data. In addition, AI must be robust to outside attacks. Pew Research found that 79% of Americans are concerned about the way their data is being used by companies.[17]

Organizations have a responsibility to use AI ethically as the technology matures. We should be fully compliant with the applicable portions of EU's GDPR and any comparable regulations in other countries, to make sure users understand that AI is working in their best interests. AI should be used to amplify our privacy, rather than undermine it.

"81% of consumers say they became more concerned over the prior year with how companies use their data, and 75% percent are now less likely to trust organizations with their personal information."

– Advancing AI ethics beyond compliance[18]

Recommended actions to take

01
Users should always maintain control over what data is being used and in what context. They can deny access to personal data that they may find compromising or unfit for an AI system to know or use.

02
Users' data should be protected from theft, misuse, or data corruption. AI systems must ensure privacy at every turn, not only with raw data, but with the insights gained from that data.

03
Provide full disclosure on how the personal information is being used or shared.

04
Allow users to deny service or data by having the AI system ask for permission before an interaction or providing the option during an interaction. Privacy settings and permissions should be clear, findable, and adjustable.

05
Forbid use of another company's data without permission when creating a new AI service.

06
Recognize and adhere to applicable national and international rights laws when designing for an AI system's acceptable user data access permissions.

To consider
Employ security practices including encryption, access control methodologies, and proprietary consent management modules to restrict access to authorized users and to de-identify data in accordance with user preferences.

It's your responsibility to work with your team to address any lack of these practices.

Questions for your team
What types of sensitive personal data does the AI system utilize and how will this data be protected?

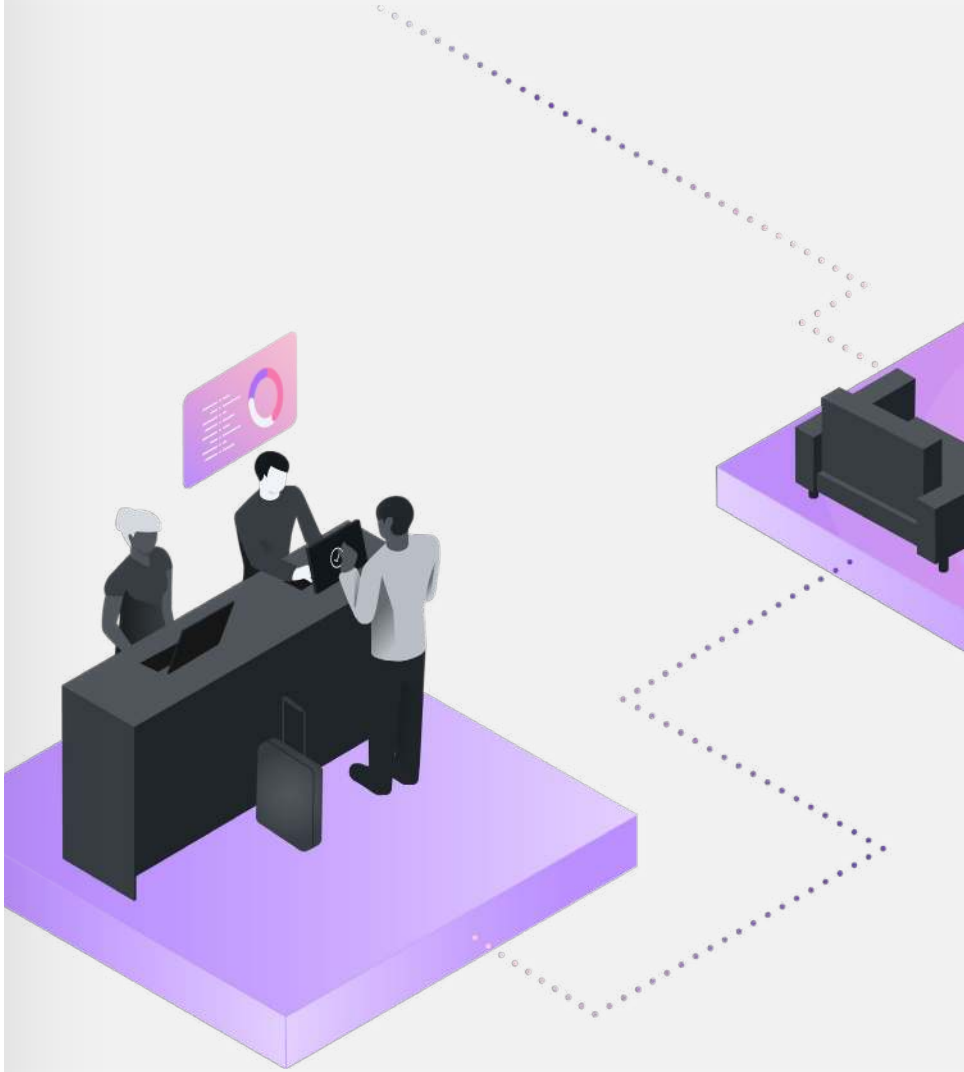
What contractual agreements are necessary for data usage and what are the local and international laws that are applicable to our AI system?

How do we create the best user experience with the minimum amount of required user data?

Running Example

– The hotel provides guests with a consent agreement to utilize the hotel's AI assistant before they begin using its services. This agreement clearly outlines to guests that the hotel does not own their data and the guests have the right to purge this data from the system at any time, even after checkout.

– During user interviews, the design researchers find that the guests feel they should be provided with a summary of the information that was acquired from them during their stay. At checkout, they can instruct the hotel to remove this information from the system if they wish.



Closing

Designers and developers of AI can help mitigate bias and disenfranchisement through these five practices.

Good intentions and general statements aren’t enough. In order to ensure our AI systems are trustworthy, we must tie our trustworthy AI pillars to our practices.

AI systems must remain flexible enough to undergo constant maintenance and improvement as ethical challenges are discovered and remediated.

By adopting the five practices covered in this document, designers and developers can become more ethically aware, mitigate biases within these systems, and instill responsibility and accountability in those who work with AI. As much of what we do related to AI is new territory for all of us, individuals and groups will need to further define criteria and metrics for evaluation to better allow for the detection and mitigation of any issues.

This is an ongoing project: we welcome and encourage feedback so the guide can develop and mature over time. We hope it contributes to the dialogue and debate about the implications of these technologies for humanity and allows designers and developers to embed ethics into the AI solutions they work on.

References

01 https://www.ibm.com/artificial-intelligence/ethics
02 https://www.ibm.com/cloud/learn/ai-ethics
03 https://www.ibm.com/design/thinking/page/badges/ai
04 https://research.ibm.com/topics/trustworthy-ai
05 https://www.weforum.org/whitepapers/responsible-use-of-technology-the-ibm-case-study/
06 https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx
07 https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
08 https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness
09 https://newsroom.ibm.com/Principles-and-Practices-for-Building-More-Trustworthy-AI
10 https://www.ibm.com/design/research/
11 https://www.ethicscanvas.org/
12 https://www.ibm.com/trust
13 https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights
14 https://newsroom.ibm.com/Principles-and-Practices-for-Building-More-Trustworthy-AI
15 https://www.ibm.com/artificial-intelligence/ai-ethics-focus-areas
16 https://gdpr-info.eu/
17 "Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information." Pew Research Center, Washington, D.C. (2019) https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/
18 https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

Download the latest version here:
ibm.biz/everydayethics