Everyday Ethics for Artificial Intelligence

A practical guide for designers & developers

IBM

September 2018

Adam Cutler: IBM Distinguished Designer
Milena Pribić: Designer
Lawrence Humphrey: Designer

Acknowledgements

Francesca Rossi - AI Ethics Global Leader, Distinguished Research Staff Member, IBM Research
Anna Sekaran - IBM Program Lead, Cognitive Computing, Academic Initiatives & Public Affairs
Jim Spohrer - Director, Cognitive OpenTech at IBM
Mike Monteiro - Co-founder, Design Director, Mule Design
Zack Causey: Designer/Illustrator

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

This report is intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. IBM shall not be responsible for any loss whatsoever sustained by any organization or person who relies on this publication.

Table of Contents

Using this Document

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

– Frank Lloyd Wright

This document represents the beginning of a conversation defining Everyday Ethics for AI. Ethics must be embedded into the design and development process from the very beginning of AI creation.

This is meant to stimulate ideas and provoke thought. The idea here is to **start simple and iterate.**

Rather than strive for perfection first, we're releasing this to allow all who read and use this to comment, critique and participate in all future iterations. **So 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 intentionally put these ideas into practice.**

As you work with your team and others, please share this guide with them.

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

Adam Cutler

IBM Distinguished Designer,

Artificial Intelligence Design

Milena Pribić

IBM Designer,

Artificial Intelligence Design

Lawrence Humphrey

IBM Designer,

Artificial Intelligence Design

Stay up to date!

Click here to download the latest version.

Introduction

Ethical decision-making is not 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.

Artificial intelligence technology 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 tech-centric focus that solely revolves around improving the capabilities of an intelligent system doesn't sufficiently consider human needs.

An ethical, human-centric AI must be designed and developed in a manner that is aligned with the values and ethical principles of a society or the community it affects.

Ethics is based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues (Markkula Center for Applied Ethics [1]).

This guide provides discussion points concerning:

a) specific virtues that AI systems should possess;

b) guidance for designers and developers building and training AI.

To create and foster trust between humans and machines, you must understand the ethical resources and standards available for reference during the designing, building, and maintenance of AI.

The large-scale focus on AI ethics by groups like the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[2], which will be referenced throughout this document, should be mirrored in businesses and working groups of all sizes.

The criteria and metrics for ethical 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.

Five Areas of Ethical Focus

01. Accountability

02. Value Alignment

03. Explainability

04. Fairness

05. User Data Rights

It's our collective responsibility to understand and evolve these ethical focus areas as AI capabilities increase over time. These focal areas provide an intentional framework for establishing an ethical foundation for building and using AI systems.

Designers and developers of AI may want to delve deeper, exploring topics covered in an IEEE course called "Artificial Intelligence and Ethics in Design[3]," including (1) Responsible Innovation in the Age of AI: Philosophical foundation, and companies using AI for profits and societal purpose; (2) The Economic Advantage of Ethical Design for Business: Intelligent systems, ethics, and government policies; (3) Values by Design in the Algorithmic Era: Identify, analyze, and practice moral, societal, and legal values , (4) The Nature of Nudging: AI ability to influence people can be used for good or bad; (5) Data Protection and Data Safety: General Data Protection Regulation and the importance of data in building and maintaining AI systems.

In addition, an IBM Research team has proposed a **Supplier's Declaration of Conformity**[4] (SDoC, or factsheet, for short) to be completed and voluntarily released by AI service developers and providers to increase the transparency of their services and engender trust in them. Like nutrition labels for foods or information sheets for appliances, factsheets for AI services would provide information about the product's important characteristics. We hope to further develop this proposal alongside our focus areas in order to usher in the era of trusted AI systems and bootstrapping their broader adoption.

Running Example

A hotel chain wants to embed artificial intelligence into an in-room virtual assistant/concierge to augment and personalize their users' stay. 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.

01.

Accountability

AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes.

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.

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.

Recommended actions to take:

1. 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.**

2. 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.

3. 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 encourage iteration.

4. Adhere to your company's business conduct guidelines. Also, understand national and international laws, regulations, and guidelines[5] that your AI may have to work within. You can find other related resources in the IEEE Ethically Aligned Design document[6].

> “Nearly 50% of the surveyed developers believe that the humans creating AI should be responsible for considering the ramifications of the technology. Not the bosses. Not the middle managers. The coders.”

- Mark Wilson, Fast Company[7] on Stack Overflow's Developer Survey Results 2018[8]

To consider:

1. Understand the workings of your AI even if you’re not personally developing and monitoring its algorithms.
2. Refer to secondary research by sociologists, linguists, behaviorists, and other professionals to understand ethical issues in a holistic context.

Questions to ask of your team:

1. How does accountability change according to the levels of user influence over an AI system?
2. Is the AI to be embedded in a human decision-making process, is it making decisions on its own, or is it a hybrid?
3. How will our team keep records of our process?
4. How do we keep track of ethical design choices and considerations after the launch of the AI?
5. Will others new to our effort be able to understand our records?

Accountability 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 a 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 at any point during their stay.

02.

Value Alignment

AI should be designed to align with the norms and values of your user group in mind.

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 job 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. 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.

Recommended actions to take:

1. 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.
2. Work with design researchers to understand and reflect your users' values. You can find out more about this process here[9].
3. Consider mapping out your understanding of your users' values and aligning the AI's actions accordingly with an Ethics Canvas.[10] Values will be specific to certain use cases and affected communities. Alignment will allow users to better understand your AI's actions and intents.

"If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms.

A necessary step in enabling machines to do so is to identify these norms. But whose norms?"

– The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [11]

To consider:

1. If you need somewhere to start, consider IBM's Standards of Corporate Responsibility[12] or use your company's standards documentation.

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

3. 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 to ask of your team:

1. Which group's values are expressed by our AI and why?

2. How do we agree on which values to consider as a team? (For more reading on moral alignment, check here.[13])

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

Value Alignment 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 hotel 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 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 will be able to speak in guests' respective languages and respect applicable customs.

03.

Explainability

AI should be designed for humans to easily perceive, detect, and understand its decision process.

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 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 the AI's conclusions and recommendations. Your users should always be aware that they are interacting with an AI. Good design does not sacrifice transparency in creating a seamless experience. **Imperceptible AI is not ethical AI.**

Recommended actions to take:

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

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

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

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

> "IBM supports transparency and data governance policies that will ensure people understand how an AI system came to a given conclusion or recommendation. Companies must be able to explain what went into their algorithm's recommendations. If they can't, then their systems shouldn't be on the market."
>
> – Data Responsibility at IBM [15]

To consider:

1. Explainability is needed to build public confidence in disruptive technology, to promote safer practices, and to facilitate broader societal adoption.
2. 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.*
3. 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 to ask of your team:

1. How do we build explainability into our experience without detracting from user experience or distracting from the task at hand?
2. Do certain processes or pieces of information need to be hidden from users for security or IP reasons? How is this explained to users?
3. Which segments of our AI decision processes can be articulated for users in an easily digestible and explainable fashion?

Explainability 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 to show how the AI makes its recommendations and suggestions.
- A researcher on the team, through interviews with hotel guests, understands that the guests want a way to opt into having their personal information stored. The team enables a way for the AI to provide guests (through voice or graphic UI) with options and the ability for the system to gather pieces of information with consent.
- With permission, the AI 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.

04.

Fairness

AI must be designed to minimize bias 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. It is the role of a responsible team to minimize algorithmic bias through ongoing research and data collection which is representative of a diverse population.

Recommended actions to take:

1. 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.

2. 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 page 26)

3. Instill a feedback mechanism or open dialogue with users to raise awareness of user-identified biases or issues. *e.g., Woebot*[17] *asks "Let me know what you think," after suggesting a link.*

> “By progressing new ethical frameworks for AI and thinking critically about the quality of our datasets and how humans perceive and work with AI, we can accelerate the [AI] field in a way that will benefit everyone. IBM believes that [AI] actually holds the keys to mitigating bias out of AI systems – and offers an unprecedented opportunity to shed light on the existing biases we hold as humans.“

– Bias in AI: How we Build Fair AI Systems and Less-Biased Humans [18]

To consider:

1. 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.

2. Your AI 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.

Questions to ask of your team:

1. How can we identify and audit unintentional biases that we run into during the design and development of our AI?

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

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

Fairness example

- After sitting down with members of the hotel’s global management, the team uncovers that diversity and inclusiveness are important elements to the hotel’s values. As a result, the team ensures that the data collected about a user's race, gender, etc. in combination with their usage of the 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 proceeds to take the time to train the model further 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 patter 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.

05.

User Data Rights

AI must be designed to protect user data and preserve the user's power over access and uses.

It is your team's responsibility to keep users empowered with control over their interactions.

Pew Research19 recently found that being in control of our own information is "very important" to 74% of Americans. The European Commission[20] found that 71% of EU citizens find it unacceptable for companies to share information about them without their permission. These percentages will rise as AI is further used to either amplify our privacy or undermine it. Your company should be fully compliant with the applicable portions of EU's General Data Protection Regulation[21] and any comparable regulations in other countries, to make sure users understand that AI is working in their best interests.

Recommended actions to take:

1. 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 to know or use.
2. Allow users to deny service or data by having the AI ask for permission before an interaction or providing the option during an interaction. Privacy settings and permissions should be clear, findable, and adjustable.
3. Provide full disclosure on how the personal information is being used or shared.
4. Users' data should be protected from theft, misuse, or data corruption.
5. Forbid use of another company's data without permission when creating a new AI service.
6. Recognize and adhere to applicable national and international rights laws[22] when designing for an AI's acceptable user data access permissions.

> “Individuals require mechanisms to help curate their unique identity and personal data in conjunction with policies and practices that make them explicitly aware of consequences resulting from the bundling or resale of their personal information.”
>
> – The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[23]

To consider:

1. 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.

2. It is your responsibility to work with your team to address any lack of these practices.

Questions to ask of your team:

1. What types of sensitive personal data does the AI utilize and how will this data be protected?

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

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

Data Rights example

- The hotel provides guests with a consent agreement to utilize the AI hotel assistant before they begin using the AI’s services. This agreement clearly outlines to guests that the hotel does not own their data and they 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 by practicing within these five areas of ethical considerations. AI systems must remain flexible enough to undergo constant maintenance and improvement as ethical challenges are discovered and remediated.

By adopting and practicing the five focal areas 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 artificial intelligence 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

1. https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/
2. https://ethicsinaction.ieee.org/
3. theinstitute.ieee.org/resources/products-and-services/new-ieee-courses-on-ethics-and-ai-and-autonomous-systems
4. https://arxiv.org/abs/1808.07261
5. http://www.ohchr.org/EN/pages/home.aspx
6. https://ethicsinaction.ieee.org
7. https://www.fastcodesign.com/90164226/what-developers-really-think-about-ai-and-bias
8. https://insights.stackoverflow.com/survey/2018/
9. https://www.ibm.com/design/research/
10. https://www.ethicscanvas.org
11. https://ethicsinaction.ieee.org
12. https://www.ibm.com/ibm/responsibility/2015/at_a_glance/our_approach.html
13. http://faculty.mtsac.edu/cmcgruder/moraljudgements.html
14. https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html
15. https://www.ibm.com/blogs/policy/dataresponsibility-at-ibm/
16. https://martechtoday.com/guide/gdpr-the-general-data-protection-regulation
17. https://woebot.io
18. https://www.ibm.com/blogs/policy/bias-in-ai/
19. www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/
20. https://ec.europa.eu/digital-single-market/en/news/eprivacy-consultations-show-confidentiality-communications-and-challenge-new-technologies-are
21. https://www.eugdpr.org/
22. http:// www.ohchr.org/EN/ProfessionalInterest/Pages/InternationalLaw.aspx
23. https://ethicsinaction.ieee.org/