Aligning Human Values and Trust in AI Systems

Why AI Designers and Users Should Care

AI is inherently value-laden. It is a common misconception that AI can make purely objective decisions. The reality is that humans encode their value systems into AI, starting with the data and abstracted into the inference model. This data is simply an aggregation of societal values collected, while the small group of designers decide to orient the model toward a certain maximization criteria.

AI designers must be aware of their own values systems - along with the ones that are built into the training data. As the experts building the technology, they would be in the best position to effectively understand and abstract the underlying values of the system to propose its best general use cases.Thus, they hold the responsibility to communicate these values to the users to build trust with the autonomous system.

Users of AI systems must understand the value systems that lie underneath the hood of these autonomous systems - especially as they delegate more and more decision making to autonomous agents. As industry stakeholders, they can act as a conscious check to the system in their daily use. They must understand how AI can supplement their human knowledge, and not become entirely dependent on it as an "objective" and "perfect" decision-making system.

Intro into Values and Trust

Decision making is more than a purely rational task. While humans can attempt to quantify decisions through economic or mathematical modeling, there will always be some assumption that oversimplifies the process, leaving a dimension unaccounted for.

Values and trust are two dimensions that are often impossible - or at least very difficult - to fully account for in economic or mathematical models. AI is a machine that can only understand the quantifiable. It takes in data and performs highly effective inferences on it. So why would AI ever be talked about as being "untrustworthy" or having "value" based decisions?

To understand what is meant by values, it is helpful to turn to the definitions made by Helen Longino, a leading figure in the philosophy of science. In her book *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry,* Longino divides values into constitutive and contextual values.

Constitutive values drive the “source of the rules determining what constitutes acceptable scientific practice or scientific method” (Longino 4). These can include ideas such as simplicity, generalizability, and accuracy of a theory along with the replicable methods of the research itself.

On the other hand, contextual values encompass the “personal, social, and cultural values, those group or individual preferences about what ought to be” as they belong to the “social and cultural environment in which science is done” (Longino 4). The broader social and cultural environment includes the moral values or ingrained infrastructure of a society which can drive the scientific progression along a certain route.

In the development of AI, the constitutive values of generalizability and accuracy are often emphasized. Autonomous systems that are generalizable and accurate are seens as an ideal. Yet, the mixture of contextual values into these systems are often overlooked, or seen as something to be avoided.

In the scope of this paper, contextual values will be the focus. AI’s decisions lie in the middle of an equation that is surrounded by human elements on both sides. Contextual values create the system - they are embedded through the societal data and the system designers. These values also surround its application - the contextual circumstances that generate the quantifiable metrics that the AI infers its decisions on.

For an AI system to be trustworthy, IBM Research argues that “we need to know that it is fair, that it’s reliable and can be accounted for, and that it will cause no harm” (Trustworthy AI). To do this, we “need assurances” that the AI cannot be tampered with, while also having the ability to “look inside AI systems” (Trustworthy AI) to understand how it came to its decision.

Both values and trust transcend quantitative metrics. There is a human element to them that cannot be fully grasped or understood. But that does not mean they completely escape quantification. In fact, they are often still embedded - yet as an unseen and unpredictable force.

In this paper, research will be presented that explores how values and trust can be ethically built into AI systems. The first paper argues that AI models must be accepted as value-laden - and explores how we must go about properly integrating contextual values into our AI systems (which has been a focal point in the “philosophy of science” field). The second paper explores how these values may be communicated and transmitted to the user - so that trust can be properly established between the system and its user.

Projects Discussed in White Paper

- "Beyond Bias and Compliance: Towards Individual Agency and Plurality of Ethics in AI," Thomas Krendl Gilbert, Megan Anne Brozek, Andrew Brozek.
- "AIX Design Framework with Character Development for Ethical AI," Sudha Jamthe, Charles Ikem

Beyond Bias and Compliance

In "Beyond Bias and Compliance," Gilbert, M. Brozek and A. Brozek argue that approaches to AI ethics must look beyond bias and compliance. The authors argue that the term "bias" is often vague and unclear - falling into a category that also includes terms like "fairness" and "responsibility." Additionally, the pursuit of compliance frames ethics as a "legal issue" that is insufficient for the field. They contend that "neither of these ideas fully encompass ethics: using moral principles to decide how to act in a particular situation" (Gilbert et al 3).

The authors evaluate the current approaches of existing companies in the AI ethics space, asserting that none of these approaches are value-agnostic. Such approaches include the following:

Approach	Description
Frameworks and Governance	AI ethics companies help ensure that teams are compliant with regulations in the industry.
AI Development Tools	Metrics are collected to ensure that the AI system is performing as expected.
Ethics Research Initiatives	Bringing researchers and industry professionals together through programs that strive to change the people involved in AI development.
AI Alignment	Companies and organizations that build general intelligent systems that are "safe" and aligned with human values.
Civil Society Advocates	Organizations that identify and diagnose AI that causes harm across human populations.
Open Source Tools	Products that have public collaboration, with the hope that increased contributions will align community values to the AI system.

All of these approaches require human based input, mixing the values of the participant into the development of the autonomous system. Gilbert et al argue that most of these solutions do not even consist of much actual ethics. Instead, they approach the problem of "aligning machines with ethical values from a psychological, sociological, or legal perspective" (Gilbert et al 7).

Approaching AI ethics from the lens of compliance reduces the discussion to "*what is permissible*, rather than *what is good*" (Gilbert et al 7). Ethics is a much broader field than this and requires self-reflection and deliberation throughout society.

This problem persists when ethics is approached as a technical problem to be solved - a downfall of the engineering mindset to "ensure standardized, repeatable, and measurable algorithms" (Gilbert et al 8). These approaches tend to facilitate discussions with the words "fairness", "responsibility", and "bias" and oversimplify the deeper philosophical questions that underlie the goal metrics.

Others approach ethics as a purely theoretical manner - and often concern themselves with the long-term AI threats such as the existential. This approach is detached and largely inapplicable to the autonomous systems that are being deployed at scale today, and "many research questions become speculative and unanswerable in the current paradigm" (Gilbert et al 9).

Finally, there are a few that approach AI ethics with an alternative framing (such as Hugging Face or Holistic AI), but these companies focus on the model parameters rather than the training data. Contextual values are built into the data itself - and this data is the ground truth for AI. Altering the way that the AI interprets what it is fed can only do little to mitigate the deeply ingrained values in its ground truth.

The Daios Method

Because of this, a new method of approaching AI ethics is needed - one that recognizes the inherent values of autonomous systems along with the philosophical reflection necessary to develop them. For these researchers, that method is the Daios method. There are three ways in which this methods diverges from existing approaches (Gilbert et al 10):

1. Data determines truth for machines.
2. Ethics is not about compliance or removing bias, but acting based on what is good.
3. Building ethical AI/ML systems first requires a neutral adjudicator to identify ethical values within those systems.

The Daios method recognizes that "the way data is labeled plays an essential role in the way AI behaves, and therefore in the ethics of machines themselves" (Gilbert et al 3). This data is all the machine can know and they "understand that information to be reality" (Gilbert et al 11). The authors argue that performing inferences on this data is more of a scientific inquiry - backed by the “senses” - rather than a philosophical endeavor. Thus, machines understand their decisions to be ethical because they are unable to “distinguish between actions that are descriptive (you observe something) or normative (you say it shouldn’t happen)” (Gilbert et al 12).

The authors take on the perspective of virtue ethicists, arguing that “criteria for virtuous behavior come from people - they are not otherworldly and do not come as a universally known fact” (Gilbert et al 14). Virtue is also highly dependent on the context, meaning that the ethical way to act will be based on one’s culture or a specific situation.

The Difficulties with Building Virtue into AI

Yet, autonomous systems take the decisions out of the hands of humans who can express this virtue, “abstracting away the need for deliberation and choice by those involved with the automated process” (Gilbert et al 14). But even if humans were able to intervene, they would most likely lack the proper context that is necessary to act in a virtuous manner. This is incredibly problematic for the scalability of virtuous actions with AI systems.

Returning to previous issues of ethics being divided between the theoretical and practical, Gilbert et al argue that theory and practice must be intertwined for the development of ethical AI systems. They argue that building ethical values into AI requires “multiple domains of knowledge” and “emotional intelligence, practical skills, technical proficiency, and personal life experience” (Gilbert et al 15). It is not a purely technical or structured process, requiring the empathetic and cultural understanding of a contextual society to develop.

Additionally, ethical action requires the agency of intent. To navigate morally challenging situations, humans may need to “redraw their own moral horizons by challenging what is fundamentally at stake in a given activity” (Gilbert et al 17). Ethical decision making is not simply following rules or inferences based on past data. Instead, it requires responsibility, accountability, and intentionally applying virtuous standards as a guide.

Finally, ethical decisions must be made from the point of view of a subject and lack objectivity. The authors of this paper claim that the conception of data being objective and independent of theory is false - it is just a “popular conception” (Gilbert et al 17). Any analysis of historical data must be interpreted through a lens of some type - and ethical decisions are bound to be interpreted as individuals will take on their own lenses.

The Solution

Thus, Gilbert et al propose an AI ethics solution that strives to empower virtue and ethical decisions in the user, going beyond the standard approaches oriented toward bias and compliance. Their Daios platform "teaches machines morality by co-creating algorithms with end users" (Gilbert et al 3), allowing the product values to be designed around the user's own value system. This creates a system that meets the ethical standards of the environment in which it is operating.

This approach allows for the "plurality of values and the freedom of individual expression" (Gilbert et al 3) by building a system that is "in tune with the values of the users…as opposed to in alignment with those who build or manage the system" (Gilbert et al 20). Rather than dictating the values, the authors look to empower ethical decisions through the user's own value system and "emphasizes knowledge from the side of the user, i.e., direct moral, value-specific feedback to the AI system" (Gilbert 21).

The authors claim that AI can never be value-neutral since data is never objective. Because of this, they understand the value-ladenness of AI and strive to "make those values visible to end users, rather than implicitly encode the values of the designers of other favored stakeholders into the system" (Gilbert et al 21).

Potential Difficulties with User Values

The Daios method effectively identifies the value-ladenness of AI and the need to integrate some of the value input to the users. Delegating some of this power out of the hands of the designers is an essential step in democratizing the value systems in AI applications and making them more beneficial to wider user bases.

Additionally, the agency offered through this approach empowers users to develop the virtue required to grow as a person. As AI scales and takes away decision-making from users, it will become essential that humans will still be provided these opportunities in situations of moral worth as it is fundamental to our well-being as moral individuals.

Yet, AI technologists should also be aware of the danger of providing too much freedom with the "plurality of values" to users of AI - especially with the effectiveness of such a tool. Having too much emphasis on the contextual moral framework of the individual AI user leaves the AI susceptible to being shaped into decisions of poor moral worth. Designers must be aware of how to counteract such manipulation - a design consideration that will once again be based on their own value systems.

Overall, a feedback system between user and designers could be an important step in developing better AI systems. Such a system must be carefully designed between various stakeholders and AI experts to see how it could be scaled across wide user bases. The Daois product takes an important step in this direction - by creating a user interface that prioritizes knowledge and value input from the user into the system.

AIX Design Framework

In their research, Jamthe and Ikem focus on the dynamics between AI and human relationships in private situations. In these settings, the AI is commonly in the form of a social robot or chatbot. Due to the intimacy of these interactions, trust must be established between the system and its users for it to be accepted into their daily lives.

Social robots and chatbots are being increasingly ingrained into our society. As of today, the most famous chatbots are ChatGPT or Google Bard - where users have an ongoing conversation with an artificial agent. However, social robots may also take on an increasingly large role in society going forward that go beyond generative knowledge. One such example is the robot seal Paro, which is licensed by the US Food and Drug Administration as a Class 2 medical device "as a therapeutic robot for use with older people, particularly those with dementia" (Sharkey 1). This device is designed to "encourage nurturing behavior" and will cry and respond to stroking or hitting - like a therapy animal might.

Such robots and autonomous systems are entering intimate parts of our lives. Thus, it becomes incredibly important that we can trustfully engage with these autonomous agents. Due to AI's foundational nature as a detached and scalable inferential system, Jamthe and Ikem claim that it faces many challenges such as "understanding emotions, creating original content, biases like racial profiling and misgendering" (Jamthe et al 3). While humans can "relate to emotional nuances like body language, tone of voice, and context" (Jamthe et al 3), AI is distinctly removed from that relational context.

AI Character and Personality

Jamthe and Ikem propose the Artificial Intelligence Design (AIX) framework with a "foundation layer of character that captures fixed tenets of ethical values of the AI" (Jamthe et al 1).

This AIX framework was developed as a "process of understanding users and demystifying the needs of the input/output data and how this entire computational process is communicated at the interface" (Jamthe et al 3). With this framework, the AI design process is centered around the human user and focused on how to translate the data inferences into interfaces that are intuitive and useful.

The authors claim that the foundational layer of this design process is the AI's character - its core values designed by the UX designer centered around trust, transparency and fairness. This character informs the AI's behavior, which is then abstracted into its personality.

AIX focuses on five elements that give personality to AI. This personality can then be automated from the deeper character foundation to "ensure that the human computer interface is designed with agency to the human or AI thoughtfully by the UX Designer" (Jamthe et al 3).

Element	**Description**
Gender	Female or Male names - with an interface design that can also genderize.
Tone	Illustrates a certain age and/or relatable human behavior, such as formal, respectful, funny, or light.
Communication Style	Can be pro-corrective or post-corrective in helping a user in their choices - should be designed to grow with the relationship between the system and the user.
Autonomy	Element that guides the AI to act in an agentive capacity to automate or assist in decisions.

Through the preliminary results of their experiments, Jamthe and Ikem illustrated that "personality built into AI makes humans trust the device more" (Jamthe et al 1). The integration of this AIX design framework in Amazon's Alexa "allows users to see transparently how Alexa makes its recommendations and will increase user trust in the long run" (Jamthe et al 11).

Discussion: User Virtue and AI Personality

As more research and frameworks are established in the space of AI ethics, technologists may consider how to properly build scalable, value-laden, and trustworthy AI systems together. For instance, one could potentially combine parts of the Daios methods with the AIX Design framework to create an AI system that combines user values and trust in these systems. One could take the emphasis on the "plurality of values and the freedom of individual expression" of the Daios method, but recognize the danger in overexposing the AI's decision-making process to the moral context of the user.

With this recognition, they could build a feedback loop that establishes trust and repertoire with the user through an expression of human personality. This AI could "recognize" the moral values of the user in a personable way, pointing out where they may line up on the spectrum of societal values.

With this feedback loop, the personable - but trustworthy - AI could help discourage decisions from extreme values. Of course, this system could run into its own trouble of accepting the societal “average” values as the ethical route. However, as more frameworks arise in this space, the integration of multiple frameworks will continue to be useful in developing new ones.

Conclusion

Overall, the authors in these papers argue that the starting point in developing ethical AI systems must be with the recognition of the inherent value-ladenness of the decisions in the system. One must understand how values will infiltrate the data input, formulate the design considerations, and impact the trust on the user end.

When this recognition has been established, designers and users may go about considering what and whose values must be weighed - and at what point in the process. Then, they can focus on how they can establish the correct trust between the system and its stakeholders.

AI is not human - but it is a human tool. And the dimensions of values and trust - which supersede the quantifiable realm of AI - can only be understood and established by those who are much more than rational beings.

Works Cited

Gilbert, Thomas Krendl, et al. “Beyond Bias and Compliance: Toward Individual Agency and Plurality of Ethics in AI.” Notre Dame - IBM Tech Ethics Lab.

Jamthe, Sudha, and Charles Ikem. “AIX Design Framework with Character Development for Ethical AI.” Notre Dame - IBM Tech Ethics Lab.

Longino, Helen E. Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press, 1990.

Sharkey, Amanda, and Natalie Wood. "The Paro seal robot: demeaning or enabling." Proceedings of AISB. Vol. 36. 2014.

“Trustworthy AI.” IBM Research, 9 Feb. 2021, research.ibm.com/topics/trustworthy-ai.