3 Methodologies to Guide Ethical Research and Design

In order to create machines that enhance human wellbeing, empowerment and freedom, system design methodologies should be extended to put greater emphasis on human rights, as defined in the Universal Declaration of Human Rights, as a primary form of human values. Therefore, we strongly believe that values-aligned design methodology should become an essential focus for the modern AI/AS organization.

Values-aligned system design puts human flourishing at the center of IT development efforts. It recognizes that machines should serve humans and not the other way around. It aims to create sustainable systems that are thoroughly scrutinized for social costs and advantages that will also increase economic value for organizations by embedding human values in design.

To help achieve these goals, technologists will need to embrace transparency regarding their products to increase end user trust. The proliferation of values-based design will also require a change of current system development approaches for organizations, including a commitment to the idea that innovation should be defined by human-centricity versus speed to market.

The process of utilizing multiple ethical approaches to provably aligned end user values will provide a key competitive differentiator in the algorithmic economy by prioritizing respect for individuals above exponential growth. Progressive organizations honoring values-based design will lead the creation of standards and policies that inform end users and other stakeholders, providing conscious consent for the use of their intelligent and autonomous technology.

Section 1 – Interdisciplinary Education and Research

Integrating applied ethics into education and research to address the issues of Artificial Intelligence and Autonomous Systems (AI/AS) requires an interdisciplinary approach, bringing together humanities, social sciences, science and engineering disciplines.

Issue:

Ethics is not part of degree programs.

Background

AI engineers and design teams too often fail to discern the ethical decisions that are implicit in technical work and design choices, or alternatively, treat ethical decision-making as just another form of technical problem solving. Moreover, technologists often struggle with the imprecision and ambiguity inherent in ethical language, which cannot be readily articulated and translated into the formal languages of mathematics, and computer programming associated with algorithms and machine learning. Thus, ethical issues can easily be rendered invisible or inappropriately reduced/simplified in the context of technical practice. This originates in the fact that Engineering programs do not often require coursework, training, or practical experience in applied ethics. A methodology for bridging the need of a truly interdisciplinary and intercultural education of the intricacies of technology and its effects on human society for the engineers who develop said technologies is required especially in regard to the immediacy ethical considerations of AI/AS

Candidate Recommendation

Ethics and ethical reflection need to be a core subject for engineers and technologists beginning at University level and for all advanced degrees. By making students sensitive to ethically aligned design issues before they enter the workplace, they can implement these methodologies in a cross-disciplinary fashion in their jobs. It is also important that these courses not be contained solely within an ethics or philosophy department but infused throughout arts, humanities and technology programs. Human values transcend all academic areas of focus.

We also recommend establishing an intercultural and interdisciplinary curriculum that is informed by ethicists, scientists, philosophers, psychologists, engineers and subject matter experts from a variety of cultural backgrounds that can be used to inform and teach aspiring engineers (post-secondary) about the relevance

3 Methodologies to Guide Ethical Research and Design

and impact of their decisions in designing AI/AS technologies. Even more critical is the priority to introduce a methodology for bridging the need for a truly interdisciplinary and intercultural education of the intricacies of technology into primary and secondary education programs. These courses should be part of the technical training and engineering development methodologies so that ethics becomes naturally part of the design process.

Further Resources

- A good example of such cross pollination can be found in the work and workshops organized by Ben Zevenbergen and Corinne Cath of the Oxford Internet Institute. The following workshop outcomes paper addresses some ethical issues in engineering from a multi-disciplinary perspective: Philosophy Meets Internet Engineering: Ethics in Networked Systems Research.
- The White House report on 'Preparing for the Future of AI' makes several recommendations on how to ensure that AI practitioners are aware of ethical issues by providing them with ethical training.
- The French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA) recommends including ethics classes in doctoral degree.
- Companies should also be encouraged to mandate consideration of ethics at the pre-product design stage, as was done by Lucid AI.

Issue:

We need models for interdisciplinary and intercultural education to account for the distinct issues of AI/AS.

Background

Not enough models exist for bringing engineers and designers in contact with ethicists and social scientists, both in academia and industry, so that meaningful interdisciplinary collaboration can shape the future of technological innovation.

Candidate Recommendation

This issue, to a large degree, relates to funding models, which limit cross-pollination between disciplines (see below). To help bridge this gap, more networking and collaboration between ethicists and technologists needs to happen in order to do the "translation work" between the worlds of investigating the social implications of technology and its actual design. Even if reasoning methods and models may differ across disciplines, sharing actual experience and knowhow is central to familiarize technologists with ethical approaches in other disciplines (e.g., medicine, architecture). Global professional organizations should devote specific access to resources (websites, MOOCS etc.) for sharing experience and methodologies.

3 Methodologies to Guide Ethical Research and Design

Further Resources

- Value Sensitive Design as described by Batya Friedman as well as Value-based Design as proposed by Sarah Spiekermann, both foresee the integration of value analysis into system design. Values are identified by senior executives and innovation team members; potentially supported by a Chief Officer devoted to this task. Then the identified values are conceptually analyzed and broken down to identify ways of system integration. Both approaches can be studied in more detail in Sarah Spiekermann's book, Ethical IT Innovation: A Value-Based System Design Approach.
- The methodology developed by the Internet Research Task Force's Human Rights Protocol Research Group (HRPC) is another example of a relevant methodology. Their guidelines provide us with an example of how human values, ethical or otherwise, relate and can be translated to Internet technology. Their website details how these values can be used in technology (both in language and in process) to fit into the Internet Engineering Task Force/ Internet Research Task Force (IETF/IRTF) engineering processes. In short, relevant values are identified on the basis of the Universal Declaration of Human Rights. These different rights are broken down into their various components and then matched to technical concepts in the process of an Internet protocol design. By combining the different technical concepts as they match different human rights components - protocol designers can approximate human rights through their work.

Issue:

The need to differentiate culturally distinctive values embedded in AI design.

Background

A responsible approach to embedded values (both as bias and as value by design) in ICTs, algorithms and autonomous systems will need to differentiate between culturally distinctive values (i.e. how do different cultures view privacy, or do they at all? And how do these differing presumptions of privacy inform engineers and technologists and the technologies designed by them?). Without falling into ethical relativism, it is critical in our international IEEE Global Initiative to avoid only considering western influenced ethical foundations. Other cultural ethical/moral, religious, corporate and political traditions need to be addressed, as they also inform and bias ICTs and autonomous systems.

Candidate Recommendation

Establish a leading role for Intercultural Information Ethics[xvi] (IIE) practitioners in value-by-design ethics committees informing technologists, policy makers and engineers. Clearly demonstrate through examples how cultural bias informs not only information flows and information systems but also algorithmic decision-making and value by design.

3 Methodologies to Guide Ethical Research and Design

Further Resources

- The work of David, et al. (2006) and Bielby (2015) has been guiding in this field "Cultural values, attitudes, and behaviors prominently influence how a given group of people views, understands, processes, communicates, and manages data, information, and knowledge."

- Pauleen, David J., et al. "Cultural Bias in Information Systems Research and Practice: Are You Coming From the Same Place I Am?." *Communications of the Association for Information Systems* 17.1 (2006): 17.
- Bielby, Jared. "Comparative Philosophies in Intercultural Information Ethics," *Confluence: Online Journal of World Philosophies* vol. 2, no. 1, (2015): 233-253.

Section 2 – Business Practices and AI

Businesses are eager to develop and monetize AI/AS but there is little supportive structure in place for creating ethical systems and practices around its development or use.

Issue:

Lack of value-based ethical culture and practices for industry.

Background

There is a need to create value-based ethical culture and practices for the development and deployment of products based on Autonomous Systems.

Candidate Recommendation

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership and responsibility, and need to consider system deployment contexts and/or ecosystems. The institution of such cultures would accelerate the adoption of the other recommendations associated within this section focused on Business Practices.

Further Resources

- The website of the Benefit Corporations (B Corporations) provides a good overview of a range of companies that personify this type of culture.

Issue:

Lack of values-aware leadership.

Background

Technology leaders give innovation teams and engineers too little or no direction on what human values should be respected in the design of a system. The increased importance of AI/AS systems in all aspects of our wired societies further accelerates the needs for value-aware leadership in AI/AS development.

Candidate Recommendations

Chief Values Officers

Companies need to create roles for senior-level marketers, ethicists or lawyers who can pragmatically implement ethically aligned design. A precedent for this type of methodological adoption comes from Agile Marketing [xvii] whose origin began in open source and engineering circles. Once the business benefits of Agile were clearly demonstrated to senior management, marketers began to embrace these

3 Methodologies to Guide Ethical Research and Design

methodologies. In today's algorithmic economy, organizations will quickly recognize the core need to identify and build to end-user values. A precedent for this new type of leader can be found in the idea of a Chief Values Officer created by Kay Firth-Butterfield. [xviii]

However, ethical responsibility should not be delegated to chief values officers. CVOs can support the creation of ethical knowledge in companies, but in the end all members of an innovation team will need to act responsibly throughout the design process.

Embedded Industry-Wide CSR
Given the need for engineers to understand intimately the cultural context and ethical considerations of design decisions, particularly as technologies afford greater levels of power, autonomy and surveillance, corporations should make a deliberate effort to ground engineering practice in authentic cultural inquiry. By creating the exemplar guidelines to enable every corporation to set up community-centered CSR efforts, companies can dedicate specific engineering resources to local problems using technology innovation for social good.

Further Resources

- As an example to emulate for embedded industry-wide Corporate Social Responsibility CSR, we recommend the Gamechangers 500 Index.

Issue:

Lack of empowerment to raise ethical concerns.

Background

Engineers and design teams are neither socialized nor empowered to raise ethical concerns regarding their designs, or design specifications, within their organizations. Considering the widespread use of AI/AS and the unique ethical questions it raises, these need to be identified and addressed from their inception.

Candidate Recommendation

Code of Conduct
In a paradigm that more fully recognizes and builds to human values, employees should be empowered to raise concerns around these issues in day to day professional practice, not just in extreme emergency circumstances such as whistleblowing. New organizational processes need to be implemented within organizations that broaden the scope around professional ethics and design as AI/AS has raised issues that do not fit the existing paradigms. New categories of considerations around these issues need to be accommodated as AI/AS have accelerated the need for new forms of Code of Conducts, so individuals feel proactively empowered to share their insights and concerns in an atmosphere of trust.

3 Methodologies to Guide Ethical Research and Design

Example: The British Computer Society (BCS) [xix] code of conduct holds that individuals have to: "a) have due regard for public health, privacy, security and wellbeing of others and the environment. b) have due regard for the legitimate rights of Third Parties*. c) conduct your professional activities without discrimination on the grounds of sex, sexual orientation, marital status, nationality, color, race, ethnic origin, religion, age or disability, or of any other condition or requirement. d) promote equal access to the benefits of IT and seek to promote the inclusion of all sectors in society wherever opportunities arise."

Further Resources

- The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights mitigates the issue surrounding the lack of empowerment to raise ethical concerns by suggesting that companies can implement measures that emphasize 'responsibility-by-design'. This term refers to solutions where the in-house working methods ensure that engineers have thought through the potential impact of their technology, where a responsible attitude to design is built into the workflow.

Issue:

Lack of ownership or responsibility from tech community.

Background

There is a divergence between the values the technology community sees as its responsibility in regards to AI/AS, and the broader set of social concerns raised by the public, legal, and social science communities.

The current makeup of most organizations has clear delineations between engineering, legal, and marketing arenas. Technologists feel responsible for safety issues regarding their work, but often refer larger social issues to other areas of their organization. Adherence to professional ethics is influenced by corporate values and may reflect management and corporate culture.

An organization may avoid using the word ethics, which then causes difficulties in applying generally agreed ethical standards. It is also understood that in technology or work contexts, "ethics" typically refers to a code of ethics regarding professional procedures (although codes of ethics often refer to values-driven design). Evolving language in this context is especially important as ethics regarding professional conduct often implies moral issues such as integrity or the lack thereof (in the case of whistleblowing, for instance).

Candidate Recommendations

Multidisciplinary ethics committees in engineering sciences should be generalized, and standards should be defined for how these committees operate, starting at a national level, then moving to international standards. Ethical Review Boards need to exist and to have the appropriate composition and use relevant criteria, and consider both research ethics and product

3 Methodologies to Guide Ethical Research and Design

ethics at the appropriate levels of advancement of research and development. They are not a silver bullet for all ethical conundrums, but can and should examine justifications of research or industrial projects in terms of ethical consequences. This is particularly important in the case of AI/AS as this technology is often deployed across many different sectors, politics, health care, transport, national security, the economy etc. Bringing together a multidisciplinary and diverse group of individuals will ensure that all the potential ethical issues are covered.

Further Resources

- Evolving the IRB: Building Robust Review for Industry Research by Molly Jackman of Facebook explains the differences between top down and bottom up approaches to the implementation of ethics within an organization.
- The article by van der Kloot Meijburg and ter Meulen gives a good overview of some of the issues involved in 'developing standards for institutional ethics committees'. It focuses specifically on health care institutions in the Netherlands, but the general lessons draw can also be applied to Ethical Review Boards.
- Examples of organization dealing with trade-offs (or "value trade offs") involved in the examination of the fairness of an algorithm to a specific end user population can for instance be found in the security considerations of the Internet Engineering Task Force (IETF).

Issue:

Need to include stakeholders for best context of AI/AS.

Background

Stakeholders or practitioners who will be working alongside AI and robotics technology have both interests to account for and, more importantly, insights to incorporate.

Candidate Recommendations

The interface between AI and practitioners has started to gain broader attention, e.g. IBM showing doctors using Watson,[xx] but there are many other contexts (esp. healthcare) where there may be different levels of involvement with the technology. We should recognize that, for example, occupational therapists and their assistants may have on-the-ground expertise in working with a patient, who themselves might be the "end user" of a robot or social AI technology. Their social and practical wisdom should be built upon rather than circumvented or replaced (as the dichotomy is usually framed to journalistic treatment). Technologists need to have that feedback, especially as it is not just academically oriented language about ethics but often a matter of crucial design detail gained by experience (form, sound, space, dialogue concepts).

Section 3 – Lack of Transparency

Lack of transparency about the AI/AS manufacturing process presents a challenge to ethical implementation and oversight.

Issue:

Poor documentation hinders ethical design.

Background

The limitations and assumptions of a system are often not properly documented. Oftentimes it is even unclear what data is processed or how.

Candidate Recommendations

Software engineers should be required to document all of their systems and related data flows, their performance and limitations and risks. Ethical values that have been prominent in the engineering processes should also be explicitly presented as well as empirical evidence of compliance and methodology used, such as data used to train the system, algorithms and components used and results of behavior monitoring. Criteria for such documentation could be: auditability, accessibility, meaningfulness, readability.

Further Resources

- The NATO Cybersecurity Centre for excellence (CCDCOE), addressed indicators of transparency along these lines.
- The Ethics of Information Transparency, Luciano Floridi.

Issue:

Inconsistent or lacking oversight for algorithms.

Background

The algorithms behind intelligent or autonomous systems are not subject to consistent oversight. This lack of transparency causes concern because end users have no context to know how a certain algorithm or system came to its conclusions.

Candidate Recommendations

Accountability

As touched on in the General Principles section of Ethically Aligned Design, transparency is an issue of concern. It is understood that specifics relating to algorithms or systems contain intellectual property that cannot be released to the general public. Nonetheless, standards providing oversight of the manufacturing process of intelligent and autonomous technologies need to be created to avoid harming end users.

3 Methodologies to Guide Ethical Research and Design

Michael Kearns suggests that we will need to decide to make algorithms less effective in order to achieve transparency.[xxi] Others argue that this trade-off is not necessary if we can devise new ways to ensure algorithmic accountability, for instance via the creation of an "algorithm FDA", or as suggested in a recent EU report through the creation of a regulatory body.[xxii] Although the discussion on what would be the best approach to create a standard is ongoing, the need for a standard is evident.

Policymakers are also free to restrict the scope of computational reasoning too complex to be understood in a conventional narrative or equations intelligible to humans. They may decide: if a bank can't give customers a narrative account of how it made a decision on their loan application, including the data consulted and algorithms used, then the bank cannot be eligible for (some of) the array of governmental perquisites or licenses so common in the financial field. They may even demand the use of public credit scoring models. (This is also a concern at the core of campaigns regarding lethal autonomous weapons: maybe countries should not develop killing machines powered by algorithms that evolve in unpredictable ways in response to unforeseeable stimuli).

Further Resources

- Frank Pasquale, Professor of Law at the University of Maryland, provides the following insights regarding accountability in a February, 2016 post for the Media Policy Project Blog produced by The London School of Economics and Political Science. He points out that even if machine learning processes are highly complex"*…we may still want to know what data was fed into the computational process. Presume as complex a credit scoring system as you want. I still want to know the data sets fed into it, and I don't want health data in that set—and I believe the vast majority agree with me on that. An account of the data fed into the system is not too complex for a person to understand, or for their own software to inspect. A relatively simple set of reforms could greatly increase transparency here, even if big data proxies can frustrate accountability."*

Issue:

Lack of an independent review organization.

Background

We need unaffiliated, expert opinions that provide guidance to the general public regarding automated systems and artificial intelligence. Currently, there is a gap between how AI/AS is marketed and their actual performance, or application. We need to ensure that AI/AS technology is accompanied by best use recommendations, and associated warnings. Additionally, we need to develop a certification scheme for AI/AS that ensures that the technologies have been independently assessed as being safe and ethically sound.

3 Methodologies to Guide Ethical Research and Design

For example, today it is possible for systems to download new parking intelligence to cars, and no independent reviewer establishes or characterizes boundaries or use. Or, when a companion robot like Jibo promises to watch your children, there is no organization that can issue an independent seal of approval or limitation on these devices. We need a ratings and approval system ready to serve social/automation technologies that will come online as soon as possible.

Candidate Recommendations

An independent, internationally coordinated body should be formed to oversee whether products actually meet ethical criteria, both when deployed, and considering their evolution after deployment and interaction with other products. Andrew Tutt's paper on an FDA for algorithms provides a good start. He argues that such an algorithm FDA would ensure that AI/AS develop in a way that is safe by: helping develop performance, design, and liability standards for algorithms, ensuring multi-stakeholder dialogue in the development of algorithms that are accountable and transparent, and ensure that AI/AS technology enters the market when it is deemed safe.

We also need further government funding for research into how AI/AS technologies can best be subjected to review, and how review organizations can consider both traditional health and safety issues, and ethical considerations.

Further Resources

- Tutt, Andrew. "An FDA for Algorithms." *Administrative Law Review* 67, 2016.

Issue:
Use of black-box components.

Background

Software developers regularly use 'black-box' components in their software, the functioning of which they often do not fully understand. 'Deep' machine learning processes, which are driving many advancements in autonomous systems, are a growing source of 'black-box' software. At least for the foreseeable future, AI developers will likely be unable to build systems that are guaranteed to operate exactly as intended or hoped for in every possible circumstance. Yet, the responsibility for resulting errors and harms remains with the humans that design, build, test and employ these systems.

Candidate Recommendations

When systems are built that could impact the safety or wellbeing of humans, it is not enough to just presume that a system works. Engineers must acknowledge and assess the ethical risks involved with black-box software and implement mitigation strategies where possible.

Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. To the

3 Methodologies to Guide Ethical Research and Design

degree that we can, it should be predictive, but given the nature of AI/AS systems it might need to be more retrospective and mitigation oriented.

Similar to the idea of a flight data recorder in the field of aviation, this algorithmic traceability can provide insights on what computations led to specific results ending up in questionable or dangerous behaviors. Even where such processes remain somewhat opaque, technologists should seek indirect means of validating results and detecting harms.

Software engineers should employ black-box software services or components only with extraordinary caution and ethical care, as they tend to produce results that cannot be fully inspected, validated or justified by ordinary means, and thus increase the risk of undetected or unforeseen errors, biases and harms.

Further Resources

- Pasquale, F. *The Black Box Society*. Harvard University Press, 2015.
- Another excellent resource on these issues can be found in Chava Gourarie's article, Investigating the algorithms that govern our lives (Columbia Journalism Review, April 2016). These additional recommended readings are referenced at the end of the article:
- "How big data is unfair": A layperson's guide to why big data and algorithms are inherently biased.
- "Algorithmic accountability reporting: On the investigation of black boxes": The primer on reporting on algorithms, by Nick Diakopoulos, an assistant professor at the University of Maryland who has written extensively on the intersection of journalism and algorithmic accountability.
- "Certifying and removing disparate impact": The computer scientist's guide to locating and fixing bias in algorithms computationally, by Suresh Venkatasubramanian and colleagues.
- *The Curious Journalist's Guide to Data:* Jonathan Stray's guide to thinking about data as communication, much of which applies to reporting on algorithms as well.