Methodologies to Guide Ethical Research and Design

To ensure autonomous and intelligent systems (A/IS) are aligned to benefit humanity A/IS research and design must be underpinned by ethical and legal norms as well as methods. We strongly believe that a value-based design methodology should become the essential focus for the modern A/IS organization.

Value-based system design methods put human advancement at the core of A/IS development. Such methods recognize that machines should serve humans, and not the other way around. A/IS developers should employ value-based design methods to create sustainable systems that are thoroughly scrutinized for social costs and advantages that will also increase economic value for organizations. To create A/IS that enhances human well-being and freedom, system design methodologies should also be enriched by putting greater emphasis on internationally recognized human rights, as a primary form of human values.

To help achieve these goals, researchers and technologists need to embrace transparency regarding their processes, products, values, and design practices to increase end-user and community trust. It will be essential that educational institutions inform engineering students about ethics, justice, and human rights, address ethical research and business practices surrounding the development of A/IS, and attend to the responsibility of the technology sector vis-à-vis public interest issues. The proliferation of value-based design will require a change of current system development approaches for organizations, including a commitment of research institutions to strong ethical guidelines for research, and of businesses to values that transcend narrow economic incentives.

Disclaimer: *While we have provided recommendations in this document, it should be understood these do not represent a position or the views of IEEE but the informed opinions of Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

Section 1 — Interdisciplinary Education and Research

Integrating applied ethics into education and research to address the issues of autonomous and intelligent systems (A/IS) requires an interdisciplinary approach, bringing together humanities, social sciences, science, engineering, and other disciplines.

Issue:

Inadequate integration of ethics in A/IS-related degree programs.

Background

AI engineers and design teams too often fail to thoroughly explore the ethical considerations implicit in their technical work and design choices. They tend to treat ethical decision-making as another form of technical problem solving. Although ethical challenges often have technical solutions, identifying and ameliorating those challenges requires technicians to methodically inquire about the social context of their work. 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 and simplified in the context of technical practice. This originates in the fact that many engineering programs do not sufficiently integrate coursework, training, or practical experience in applied ethics throughout their curricula; too often ethics is relegated to a stand-alone course or module that gives students little or no direct experience in ethical decision-making in engineering work. Ethics education for engineering students should be meaningful, measurable, and incorporate best practices of STEM ethics education drawn from pertinent multidisciplinary resources.

The aim of these recommendations is to prepare students for the technical training and engineering development methodologies that incorporate ethics as essential so that ethics and human rights become naturally part of the design process.

Candidate Recommendations

Ethics and ethical reflection need to be a core subject for aspiring engineers and technologists beginning at the earliest appropriate level and for all advanced degrees. By training students how to be sensitive to ethical issues in design before they enter the workplace, they can more effectively implement value-based design methodologies in the context of A/IS work.

Methodologies to Guide Ethical Research and Design

We also recommend that effective STEM ethics curricula be informed by *scientists, artists, philosophers, psychologists, legal scholars, engineers, and other subject matter experts* from a variety of cultural backgrounds to ensure that students acquire sensitivity to a diversity of robust perspectives on human flourishing. Such curricula should teach aspiring engineers, computer scientists, and statisticians about the relevance and impact of their decisions in designing A/IS technologies. Effective ethics education in STEM contexts should span primary, secondary, and post-secondary education, and include both universities and vocational training schools. Relevant accreditation bodies should reinforce this integrated approach as outlined above.

Further Resources

- Holdren, J., and M. Smith. "Preparing for the Future of Artificial Intelligence." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016. This White House report 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 programs.
- The U.S. National Science Foundation has funded extensive research on STEM ethics education best practices through the Cultivating Cultures for Ethical Science, Technology, Engineering, and Mathematics (CCE-STEM) Program, and recommends integrative approaches that incorporate ethics throughout STEM education.
- Comparing the UK, EU, and US approaches to AI and ethics: Cath, C. et al. "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach." *Science and Engineering Ethics* (2017).
- The Oxford Internet Institute (OII) organized a workshop on ethical issues in engineering. The output paper can be found here: Zevenbergen, B. et al. "Philosophy Meets Internet Engineering: Ethics in Networked Systems Research." Oxford, U.K.: Oxford Internet Institute, University of Oxford, 2015.
- Companies should also be encouraged to mandate consideration of ethics at the pre-product design stage, as was done by Lucid AI.
- There are a variety of peer-reviewed online resources collecting STEM ethics curricula, syllabi, and education modules:
  - Ethics Education Library, Illinois Institute of Technology
  - IDEESE: International Dimensions of Ethics Education in Science & Engineering, University of Massachusetts Amherst
  - National Center for Professional & Research Ethics, University of Illinois
  - Online Ethics Center, National Academy of Engineering

Methodologies to Guide Ethical Research and Design

Issue:

The need for more constructive and sustained interdisciplinary collaborations to address ethical issues concerning autonomous and intelligent systems (A/IS).

Background

Not enough institutional resources and incentive structures exist for bringing A/IS engineers and designers into sustained and constructive contact with ethicists, legal scholars, and social scientists, both in academia and industry. This contact is necessary as it can enable meaningful interdisciplinary collaboration to shape the future of technological innovation. There are currently few methodologies, shared knowledge, and lexicons that would facilitate such collaborations.

This issue, to a large degree, relates to funding models as well as the traditional mono-function culture in A/IS-related institutions and companies, which limit cross-pollination between disciplines (see below). To help bridge this gap, additional "translation work" and resource sharing (including websites and MOOCs) needs to happen among technologists and other relevant experts (e.g., in medicine, architecture, law, philosophy, psychology, cognitive science).

Candidate Recommendations

Funding models and institutional incentive structures should be reviewed and revised to prioritize projects with interdisciplinary ethics components to encourage integration of ethics into projects at all levels.

Further Resources

- Barocas, S. Course Material for Ethics and Policy in Data Science.
- Floridi, L., and M. Taddeo. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society* 374, no. 2083 (2014): 1–4. doi:10.1098/rsta.2016.0360.
- Spiekermann, S. *Ethical IT Innovation: A Value-Based System Design Approach*. Boca Raton, Florida: Auerbach Publications, 2015.
- The approach developed by the Internet Research Task Force's Human Rights Protocol Research Group (HRPC) for integrating human rights concern in technical design.

Issue:

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

Background

A responsible approach to embedded values (both as uncritical bias and as value by design) in information and communications technology (ICT), algorithms and autonomous systems will need to differentiate between culturally distinctive values (i.e., how do different cultures

Methodologies to Guide Ethical Research and Design

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 oversimplified ethical relativism, or embedding values that are antithetical to human flourishing (for example, human rights violations), it is critical that A/IS design avoids only considering monocultural influenced ethical foundations.

Candidate Recommendations

Establish a leading role for intercultural information ethics (IIE) practitioners in 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.

Further Resources

- Pauleen, D. 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, no. 17 (2006). The work of Pauleen et al. (2006) and Bielby (2015) has been guiding in this field: "Cultural values, attitudes, and behaviours prominently influence how a given group of people views, understands, processes, communicates, and manages data, information, and knowledge."
- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics," *Confluence: Online Journal of World Philosophies* 2, no. 1 (2015): 233–253.

Section 2 – Corporate Practices and A/IS

Corporations, whether for-profit or not-for-profit, are eager to develop, deploy, and monetize A/IS, but there are insufficient structures in place for creating and supporting ethical systems and practices around A/IS funding, 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 and intelligent systems (A/IS). To do so, we need to further identify and refine social processes and management strategies that facilitate values-based design in the engineering and manufacturing process.

Candidate Recommendations

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership, and responsibility, and the need to consider system deployment contexts and/or ecosystems. The institution of an ethical A/IS corporate culture would accelerate the adoption of the other recommendations 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.
- *Firms of Endearment* is a book which showcases how companies embracing values and a stakeholder approach outperform their competitors in the long run.
- The ACM Code of Ethics and Professional Ethics, which also includes various references to human well-being and human rights.

Methodologies to Guide Ethical Research and Design

Issue:

Lack of values-aware leadership.

Background

Technology leadership should give innovation teams and engineers direction regarding which human values and legal norms should be promoted in the design of an A/IS system. Cultivating an ethical corporate culture is an essential component of successful leadership in the A/IS domain.

Candidate Recommendations

Companies need to create roles for senior-level marketers, ethicists, or lawyers who can pragmatically implement ethically aligned design, both the technology and the social processes to support value-based system innovation. Companies need to ensure that their understanding of value-based system innovation is based on *de jure* and *de facto* international human rights standards.

A promising way to ensure values are on the agenda in system development is to have a Chief Values Officer (CVO), a role first suggested by Kay Firth-Butterfield, Vice-Chair, The IEEE Global Initiative and Project Head of AI and Machine Learning at the World Economic Forum. The CVO should support system innovations and engineering teams to consider values and provide them with methodological guidance on how to do so. However, ethical responsibility should not be delegated solely to CVOs. 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.

Further Resources

- United Nations, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, New York and Geneva: UN, 2011.
- Institute for Human Rights and Business (IHRB), and Shift, SectICTor Guide on Implementing the UN Guiding Principles on Business and Human Rights, 2013.
- Cath, C., and L. Floridi. "The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights." *Science and Engineering Ethics* 23, no. 2 (2017): 449–468.
- Butterfield, Kay-Firth (2017). How IEEE Aims to Instill Ethics in Artificial Intelligence Design. *The Institute*.

Methodologies to Guide Ethical Research and Design

Issue:

Lack of empowerment to raise ethical concerns.

Background

Engineers and design teams can encounter obstacles to raising ethical concerns regarding their designs or design specifications within their organizations. Corporate culture should incentivize technical staff to voice the full range of ethical questions to relevant corporate actors throughout the full product lifecycle. Because raising ethical concerns can be perceived as slowing or halting a design project, organizations need to consider how they can recognize and incentivize value-based design as an integral component of product development.

Candidate Recommendations

Employees should be empowered to raise ethical concerns in day-to-day professional practice, not just in extreme emergency circumstances such as whistleblowing. New organizational and socio-cultural processes that broaden the scope around professional ethics and design need to be implemented within organizations. New categories of considerations around these issues need to be accommodated along with new forms of Codes of Conduct, so individuals are empowered to share their insights and concerns in an atmosphere of trust.

Further Resources

- The British Computer Society (BCS) code of conduct holds that individuals have to: "a) have due regard for public health, privacy, security and well-being 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, colour, 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."
- 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 as they relate to human rights 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.

Methodologies to Guide Ethical Research and Design

Issue:

Organizations should examine their cultures to determine how to flexibly implement value-based design.

Background

Ethics is often treated as an impediment to innovation, even among those who ostensibly support ethical design practices. In industries that reward rapid innovation, it is necessary to develop design practices that integrate effectively with existing engineering workflows. Those who advocate for ethical design within a company should not be seen as innovators seeking the best ultimate outcomes for the company, end-users, and society. Leaders can facilitate that mindset by promoting an organizational structure that supports the integration of dialogue about ethics throughout product lifecycles.

A/IS design processes often present moments where ethical consequences can be highlighted. There are no universally prescribed models for this because organizations vary significantly in structure and culture. In some organizations, design team meetings may be brief and informal. In others, the meetings may be lengthy and structured. Regardless, team members should understand how to raise such questions without being perceived as impediments by peers and managers. The transitions point between discovery, prototyping, release, and revisions are natural contexts for conducting such reviews.

Iterative review processes are also advisable, in part because changes to risk profiles over time can illustrate needs or opportunities for improving the final product.

Candidate Recommendations

Companies should study their own design processes to identify moments where engineers and researchers can be encouraged to raise and resolve questions of ethics. Achieving a distributed responsibility for ethics requires that all people involved in product design are encouraged to notice and respond to ethical concerns, particularly around safety, bias, and legality. Organizations should consider how they can best encourage and accommodate lightweight deliberations among peers.

Additionally, organizations should identify points for formal review inside their product development processes. These reviews can focus on "red flags" that have been identified in advance as indicators of risk. For example, if the datasets involve minors or focus on users from protected classes then it may require additional justification or alterations to the research or development protocols.

Further Resources

- Sinclair, A. "Approaches to Organizational Culture and Ethics." *Journal of Business Ethics* 12, no. 1 (1993): 63–73.
- Chen, A. Y. S., R. B. Sawyers, and P. F. Williams. "Reinforcing Ethical Decision Making Through Corporate Culture." *Journal of Business Ethics* 16, no. 8 (1997): 855–865.

Methodologies to Guide Ethical Research and Design

- Crawford, K., and R. Calo. "There Is a Blind Spot in AI Research." *Nature* 538 (2016): 311–313.

Issue:

Lack of ownership or responsibility from the tech community.

Background

There is a divergence between the values the technology community sees as its responsibility in regards to A/IS, and the broader set of social concerns raised by the public, legal, and professional communities. The current makeup of most organizations has clear delineations among engineering, legal, and marketing arenas. Thus technologists feel responsible for safety issues regarding their work, but for larger social issues may say, "legal will handle that." In addition, in employment and management technology or work contexts, "ethics" typically refers to a code of conduct regarding professional decorum (versus a values-driven design process mentality). As such, ethics regarding professional conduct often implies moral issues such as integrity or the lack thereof (in the case of whistleblowing, for instance), but ethics in A/IS design includes broader considerations about the consequences of technologies.

Candidate Recommendations

To help integrate applied ethics regarding A/IS and in general, organizations need to choose specific language that will break down traditional biases or barriers and increase adoption of values-based design. For instance, an organization can refer to the "trade-offs" (or "value trade-offs") involved in the examination of the fairness of an algorithm to a specific end user population.

Organizations should clarify the relationship between professional ethics and applied A/IS ethics and help designers, engineers, and other company representatives discern the differences between them and where they complement each other.

Corporate ethical review boards, or comparable mechanisms, should be formed to address ethical concerns in relation to their A/IS research. Such boards should seek an appropriately diverse composition and use relevant criteria, including both research ethics and product ethics at the appropriate levels of advancement of research and development. These boards should examine justifications of research or industrial projects in terms of consequences for human flourishing.

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 approach to the implementation of ethics within an organization and describes Facebook's internal ethics review for research and development.

Methodologies to Guide Ethical Research and Design

- 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 drawn can also be applied to ethical review boards. Examples of organizations dealing with such trade-offs can for instance be found in the security considerations of the Internet Engineering Task Force (IETF).

Issue:

Need to include stakeholders for adequate ethical perspective on A/IS.

Background

The interface between AI and practitioners, as well as other stakeholders, is gaining broader attention in domains such as health care diagnostics, and there are many other contexts 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. Technologists need to have that stakeholder feedback, because beyond academically oriented language about ethics, that feedback is often about crucial design detail gained by experience (form, sound, space, dialogue concepts). There are successful models of user experience (UX design) that account for human factors which should be incorporated to A/IS design as systems are more widely deployed.

Candidate Recommendations

Account for the interests of the full range of stakeholders or practitioners who will be working alongside A/IS, incorporating their insights. Build upon, rather than circumvent or ignore, the social and practical wisdom of involved practitioners and other stakeholders.

Further Resources

- Schroeter, Ch. et al. "Realization and User Evaluation of a Companion Robot for People with Mild Cognitive Impairments." *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2013)*, Karlsruhe, Germany (2013): 1145–1151.
- Chen, T. L. et al. "Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities." *IEEE Robotics and Automation Magazine* 20, no. 1 (2013): 30–39.
- Hartson, R., and P. S. Pyla. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Waltham, MA: Elsevier, 2012.

Section 3 — Research Ethics for Development and Testing of A/IS Technologies

Issue:

Institutional ethics committees are under-resourced to address the ethics of R&D in the A/IS fields.

Background

It is unclear how research on the interface of humans and A/IS, animals and A/IS, and biological hazards will pose practical challenges for research ethical review boards. Norms, institutional controls, and risk metrics appropriate to the technology are not well established in the relevant literature and research governance infrastructure. Additionally, national and international regulations governing review of human-subjects research may explicitly or implicitly exclude A/IS research from their purview on the basis of legal technicalities or medical ethical concerns regardless of potential harms posed by the research.

Research on A/IS human-machine interaction, when it involves intervention or interaction with identifiable human participants or their data, typically falls to the governance of research ethics boards (e.g., institutional review boards). The national level and institutional resources (e.g., hospitals and universities) to govern ethical conduct of HCI, particularly within the disciplines pertinent to A/IS research, are underdeveloped. First, there is limited international or national guidance to govern this form of research. While sections of IEEE standards governing research on AI in medical devices address some of the issues related to security of AI-enabled devices, the ethics of testing those devices to bring them to market are not developed into recognized national (e.g., U.S. FDA) or international (e.g., EU EMA) policies or guidance documents. Second, the bodies that typically train individuals to be gatekeepers for the research ethics bodies (e.g., PRIM&R, SoCRA) are under-resourced in terms of expertise for A/IS development. Third, it is not clear whether there is sufficient attention paid to A/IS ethics by research ethics board members or by researchers whose projects involve the use of human participants or their identifiable data.

Research pertinent to the ethics governing research at the interface of animals and A/IS research is underdeveloped with respect to systematization for implementation by

Methodologies to Guide Ethical Research and Design

IACUC or other relevant committees. In institutions without a veterinary school, it is unclear that the organization would have the relevant resources necessary to conduct an ethical review of such research.

Research pertinent to the intersection of radiological, biological, and toxicological research (ordinarily governed under institutional biosafety committees) and A/IS research is not found often in the literature pertinent to research ethics or research governance. Beyond a limited number of pieces addressing the "dual use" or import/export requirements for A/IS in weapons development, there are no guidelines or standards governing topics ordinarily reserved for review by institutional biosafety committees, or institutional radiological safety committees, or laboratory safety committees.

Candidate Recommendations

IEEE should draw upon existing standards, empirical research, and expertise to identify priorities and develop standards for governance of A/IS research and to partner with relevant national agencies, and international organizations, when possible.

Further Resources

- Jordan, S. R. "The Innovation Imperative." *Public Management Review* 16, no. 1 (2014): 67–89.
- Schneiderman, B. "The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 48 (2016): 13538–13540.
- Metcalf, J., and K. Crawford. "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide." SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, 2016.
- Calo, R. "Consumer Subject Review Boards: A Thought Experiment." *Stanford Law Review Online* 66 (2013): 97.

Section 4 — Lack of Transparency

Lack of transparency about the A/IS 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 Recommendation

Software engineers should be required to document all of their systems and related data flows, their performance, 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, and readability.

Further Resources

- Cath, C. J. N., L. Glorioso, and M. R. Taddeo. "NATO CCD COE Workshop on 'Ethics and Policies for Cyber Warfare'" NATO Cybersecurity Centre for Excellence (CCDCOE) Report. Oxford, U.K.: Magdalen College. Addressed indicators of transparency along these lines.
- Turilli, M., and L. Floridi. "The Ethics of Information Transparency." *Ethics and Information Technology* 11, no. 2 (2009): 105–112.
- Wachter, S., B. Mittelstadt, and L. Floridi. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2, no. 6 (2017).
- Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. "Accountable Algorithms." *University of Pennsylvania Law Review* 165, no. 1 (2017): 633–705.
- Balkin, J. M., "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *UC Davis Law Review*, (2018 forthcoming).

Methodologies to Guide Ethical Research and Design

Issue:

Inconsistent or lacking oversight for algorithms.

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. These recommendations are similar to those made in committees 1 and 2, but here are used as they apply to the narrow scope of this group.

Candidate Recommendations

Accountability

As touched on in the General Principles section of Ethically Aligned Design, algorithmic 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 harm and negative consequences of the use of these technologies. Here we can look to other technical domains, such as biomedical, civil, and aerospace engineering, where commercial protections for proprietary technology are routinely and effectively balanced with the need for appropriate oversight standards and mechanisms to safeguard the public.

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.
- Ryan Calo, Associate Professor of Law at the University of Washington, wrote an excellent article that gives a detailed overview of a broad array of AI policy questions.
- In the United States, a recent court case, Armstrong, highlights the need for appropriate oversight of algorithmic decision-making, to preserve due process and other legal and ethical principles. K.W. v. Armstrong, 180 F. Supp. 3d 703 (D. Idaho 2016). In the case, a court ruled that Idaho's Department of Health and Welfare violated the rights of disabled Medicaid recipients by relying upon arbitrary and flawed algorithmic decision systems when cutting benefits, and refusing to disclose the decision bases as 'trade secrets.' See details of the case here: https://www.aclu.org/news/federal-court-rules-against-idaho-department-health-and-welfare-medicaid-class-action and a related discussion of the general risks of opaque algorithmic bureaucracies here: https://medium.com/aclu/pitfalls-of-artificial-intelligence-decisionmaking-highlighted-in-idaho-aclu-case-ec59941fb026

Methodologies to Guide Ethical Research and Design

Issue:
Lack of an independent review organization.

Background

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

For example, today it is possible for systems to download new self-parking functionality 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. We also need further government funding for research into how A/IS technologies can best be subjected to review, and how review organizations can consider both traditional health and safety issues, as well as ethical considerations.

Candidate Recommendations

An independent, internationally coordinated body should be formed to oversee whether such products actually meet ethical criteria, both when deployed, and considering their evolution after deployment and interaction with other products.

Further Resources

- Tutt, A. "An FDA for Algorithms." *Administrative Law Review* (2017): 83–123.
- Scherer, M. U. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law and Technology* 29, no. 2 (2016): 354–400.
- Desai, D. R., and J. A. Kroll. "Trust But Verify: A Guide to Algorithms and the Law." *Harvard Journal of Law and Technology* (2018 forthcoming).

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"

Methodologies to Guide Ethical Research and Design

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 Recommendation

When systems are built that could impact the safety or well-being 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.

Candidate Recommendation

Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. To the degree possible, these characterizations should be predictive, but given the nature of A/IS, they might need to be more retrospective and mitigation oriented. Such standards may include preferential adoption of effective design methodologies for building "explainable AI" (XAI) systems that can provide justifying reasons or other reliable "explanatory" data illuminating the cognitive processes leading to, and/or salient bases for, their conclusions.

Candidate Recommendation

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

Candidate Recommendation

Software engineers should employ "black-box" (opaque) 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*. Cambridge, MA: Harvard University Press, 2015.
- In the United States, in addition to similar commercial endeavors by Oracle and other companies, DARPA (Defense Advanced Research Projects Agency) recently funded a 5-year research program in explainable AI (XAI) methodologies.
- Ananny, M., and K. Crawford. (2016). "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, December 13, 2016.
- Another excellent resource on these issues can be found in Chava Gourarie's "Investigating the Algorithms That Govern Our Lives." *Columbia Journalism Review*, April 14, 2016. These recommended reads come at the end of the article:

Methodologies to Guide Ethical Research and Design

- "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 gentle guide to thinking about data as communication, much of which applies to reporting on algorithms as well.