ETHICALLY ALIGNED DESIGN

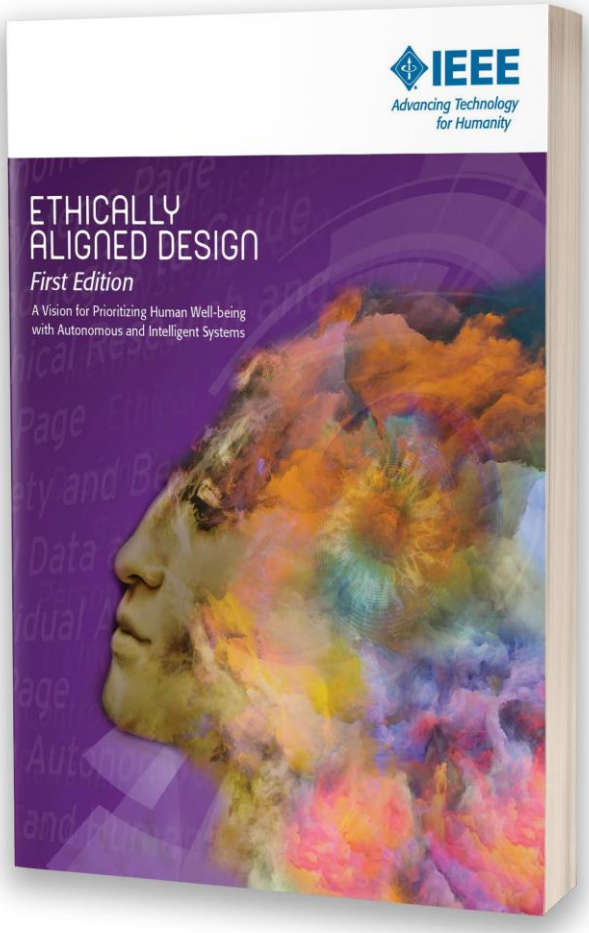
First Edition

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

IEEE

Ethically Aligned Design, First Edition

Available 25 March 2019

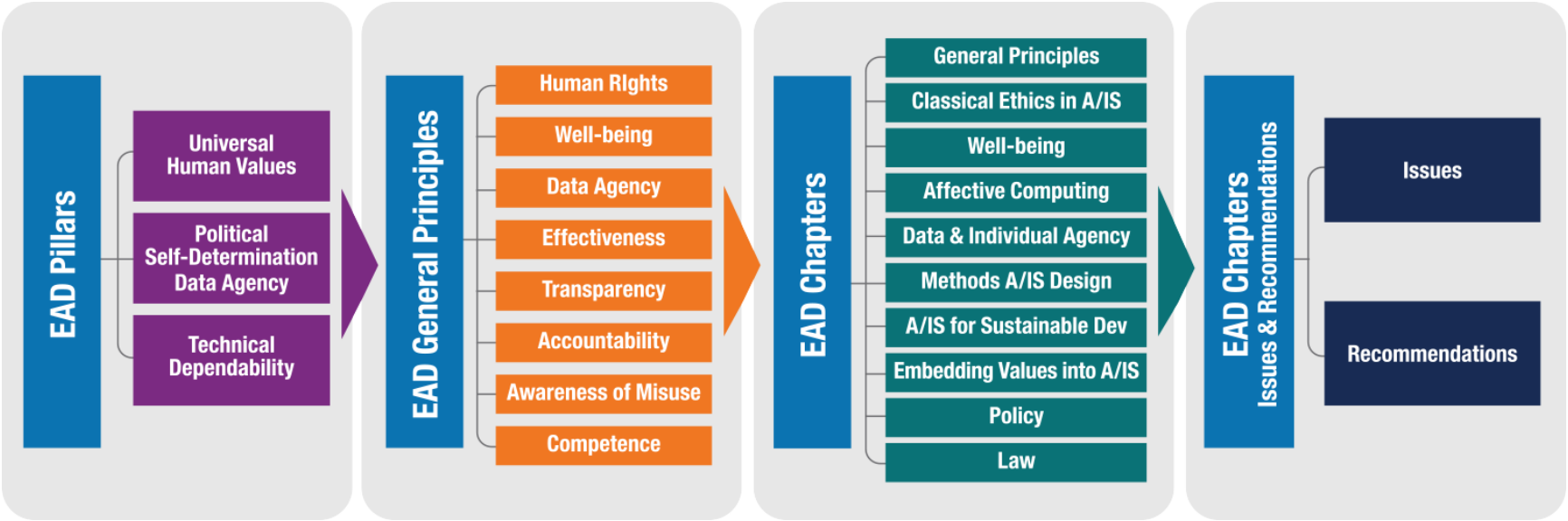


- Developed under the auspices of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
  - The IEEE Global Initiative was launched in 2016 and has grown to a community of over 2100 participants from the private sector, government, academia and civil society.
- Sets forth scientific analysis and resources, high-level principles and actionable recommendations.
- Aim is to inform the broader public and to inspire its global audience of academics, engineers, policymakers, developers and users of A/IS to take action.
- A comprehensive body of work grounded in a global, open, iterative and collaborative process
  - Providing credibility and integrity to the body of work; many/diverse voices heard.

IEEE

Ethically Aligned Design, First Edition

Conceptual Framework—From Principles to Practice



EAD Pillars

- Universal Human Values: A/IS can be an enormous force for good in society provided they are designed to respect human rights, align with human values, and holistically increase well-being while empowering as many people as possible. They should also be designed to safeguard our environment and natural resources. These values should guide policy makers as well as engineers, designers, and developers. Advances in A/IS should be in the service of all people, rather than benefiting solely small groups, a single nation, or a corporation.
- Political Self-Determination and Data Agency: A/IS—if designed and implemented properly have a great potential to nurture political freedom and democracy, in accordance with the cultural precepts of individual societies, when people have access to and control over the data constituting and representing their identity. These systems can improve government effectiveness and accountability, foster trust, and protect our private sphere, but only when people have agency over their digital identity and their data is provably protected.
- Technical Dependability: Ultimately, A/IS should deliver services that can be trusted.2 This trust means that A/IS will reliably, safely, and actively accomplish the objectives for which they were designed while advancing the human-driven values they were intended to reflect. Technologies should be monitored to ensure that their operation meets predetermined ethical objectives aligning with human values and respecting codified rights. In addition, validation and verification processes, including aspects of explain ability, should be developed that could lead to better auditability and to certification3 of A/IS.

Ethically Aligned Design, First Edition

General Principles

- **Human Rights -** A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
- **Well-being -** A/IS creators shall adopt increased human well-being as a primary success criterion for development.
- **Data Agency -** A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
- **Effectiveness -** A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
- **Transparency -** The basis of a particular A/IS decision should always be discoverable.
- **Accountability -** A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
- **Awareness of Misuse -** A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
- **Competence -** A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Human Rights

Recommendations

- To best respect human rights, society must assure the safety and security of A/IS so that they are designed and operated in a way that benefits humans. Specifically:
  - Governance frameworks, including standards and regulatory bodies, should be established to oversee processes which ensure that the use of A/IS does not infringe upon human rights, freedoms, dignity, and privacy, and which ensure traceability. This will contribute to building public trust in A/IS.
  - A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for diverse cultural norms as well as differing legal and regulatory frameworks.
  - A/IS should always be subordinate to human judgment and control.
  - For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights.

Well Being

Recommendations

- A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point.

Data Agency

Recommendations

- Organizations, including governments, should immediately explore, test, and implement technologies and policies that let individuals specify their online agent for case-by-case authorization decisions as to who can process what personal data for what purpose. For minors and those with diminished capacity to make informed decisions, current guardianship approaches should be viewed to determine their suitability in this context. The general solution to give agency to the individual is meant to anticipate and enable individuals to own and fully control autonomous and intelligent (as in capable of learning) technology that can evaluate data use requests by external parties and service providers. This technology would then provide a form of “digital sovereignty” and could issue limited and specific authorizations for processing of the individual’s personal data wherever it is held in a compatible system.

Effectiveness

Recommendations

- Creators engaged in the development of A/IS should seek to define metrics or benchmarks that will serve as valid and meaningful gauges of the effectiveness of the system in meeting its objectives, adhering to standards and remaining within risk tolerances. Creators building A/IS should ensure that the results when the defined metrics are applied are readily obtainable by all interested parties, e.g., users, safety certifiers, and regulators of the system.
- Creators of A/IS should provide guidance on how to interpret and respond to the metrics generated by the systems.
- To the extent warranted by specific circumstances, operators of A/IS should follow the guidance on measurement provided with the systems, i.e., which metrics to obtain, how and when to obtain them, how to respond to given results, and so on.
- To the extent that measurements are sample based, measurements should account for the scope of sampling error, e.g., the reporting of confidence intervals associated with the measurements. Operators should be advised how to interpret the results.
- Creators of A/IS should design their systems such that metrics on specific deployments of the system can be aggregated to provide information on the effectiveness of the system across multiple deployments. For example, in the case of autonomous vehicles, metrics should be generated both for a specific instance of a vehicle and for a fleet of many instances of the same kind of vehicle.
- In interpreting and responding to measurements, allowance should be made for variation in the specific objectives and circumstances of a given deployment of A/IS.
- To the extent possible, industry associations or other organizations, e.g., IEEE and ISO, should work toward developing standards for the measurement and reporting on the effectiveness of A/IS.

Transparency

Recommendations

- Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. The mechanisms by which transparency is provided will vary significantly, including but not limited to, the following use cases:
  - For users of care or domestic robots, a “why did-you-do-that button” which, when pressed, causes the robot to explain the action it just took.
  - For validation or certification agencies, the algorithms underlying the A/IS and how they have been verified.
  - For accident investigators, secure storage of sensor and internal state data comparable to a flight data recorder or black box.

Accountability

Recommendations

- To best address issues of responsibility and accountability:
  - Legislatures/courts should clarify responsibility, culpability, liability, and accountability for A/IS, where possible, prior to development and deployment so that manufacturers and users understand their rights and obligations.
  - Designers and developers of A/IS should remain aware of, and take into account, the diversity of existing cultural norms among the groups of users of these A/IS.
  - Multi-stakeholder ecosystems including creators, and government, civil, and commercial stakeholders, should be developed to help establish norms where they do not exist because A/IS-oriented technology and their impacts are too new. These ecosystems would include, but not be limited to, representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users. The norms can mature into best practices and law
  - Systems for registration and record-keeping should be established so that it is always possible to find out who is legally responsible for a particular A/IS. Creators, including manufacturers, along with operators, of A/IS should register key, high-level parameters, including:
    - Intended use,
    - Training data and training environment, if applicable,
    - Sensors and real world data sources,
    - Algorithms,
    - Process graphs,
    - Model features, at various levels,
    - User interfaces,
    - Actuators and outputs, and
    - Optimization goals, loss functions, and reward functions.

Awareness of Misuse

Recommendations

- Creators should be aware of methods of misuse, and they should design A/IS in ways to minimize the opportunity for these.
- Raise public awareness around the issues of potential A/IS technology misuse in an informed and measured way by:
  - Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS. For example, provide “data privacy warnings” that some smart devices will collect their users’ personal data.
  - Delivering this education in scalable and effective ways, including having experts with the greatest credibility and impact who can minimize unwarranted fear about A/IS.
  - Educating government, lawmakers, and enforcement agencies about these issues of A/IS so citizens can work collaboratively with these agencies to understand safe use of A/IS. For example, the same way police officers give public safety lectures in schools, they could provide workshops on safe use and interaction with A/IS.

Competence

Recommendations

- Creators of A/IS should specify the types and levels of knowledge necessary to understand and operate any given application of A/IS. In specifying the requisite types and levels of expertise, creators should do so for the individual components of A/IS and for the entire systems.
- Creators of A/IS should integrate safeguards against the incompetent operation of their systems. Safeguards could include issuing notifications/warnings to operators in certain conditions, limiting functionalities for different levels of operators (e.g., novice vs. advanced), system shut-down in potentially risky conditions, etc.
- Creators of A/IS should provide the parties affected by the output of A/IS with information on the role of the operator, the competencies required, and the implications of operator error. Such documentation should be accessible and understandable to both experts and the general public.
- Entities that operate A/IS should create documented policies to govern how A/IS should be operated. These policies should include the real-world applications for such A/IS, any preconditions for their effective use, who is qualified to operate them, what training is required for operators, how to measure the performance of the A/IS, and what should be expected from the A/IS. The policies should also include specification of circumstances in which it might be necessary for the operator to override the A/IS.
- Operators of A/IS should, before operating a system, make sure that they have access to the requisite competencies. The operator need not be an expert in all the pertinent domains but should have access to individuals with the requisite kinds of expertise.

For a copy of
***Ethically Aligned Design*, First Edition**
please visit:

ethicsinaction.ieee.org