IEEE is the largest technical professional organization dedicated to advancing technology for the benefit of humanity, with over 420,000 members in more than 160 countries. Through its highly cited publications, conferences, technology standards, and professional and educational activities, IEEE is the trusted voice in a wide variety of areas ranging from aerospace systems, computers, and telecommunications to biomedical engineering, electric power, and consumer electronics.

Ethically Aligned Design

THE GOAL OF THE IEEE GLOBAL INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS (“THE IEEE GLOBAL INITIATIVE”) IS THAT ETHICALLY ALIGNED DESIGN WILL PROVIDE PRAGMATIC AND DIRECTIONAL INSIGHTS AND RECOMMENDATIONS, SERVING AS A KEY REFERENCE FOR THE WORK OF TECHNOLOGISTS, EDUCATORS AND POLICYMAKERS IN THE COMING YEARS.

What is it? And Why is it important?

- With AI being used more than ever before in almost every aspect of technological advances, Ethically Aligned Design is aimed at ensuring that the use of autonomous and intelligent systems (A/IS) remain focused on being beneficial to people and the environment (eudaimonia - defines human well-being, both at the individual and collective level)
- Since AI/S is designed to operate without or less human intervention, careful consideration of human wellbeing at the point of designing will go a long way in preventing potential harm to privacy, discrimination, loss of skills, adverse economic impacts, risks to security of critical infrastructure, and possible negative long-term effects on societal well-being.

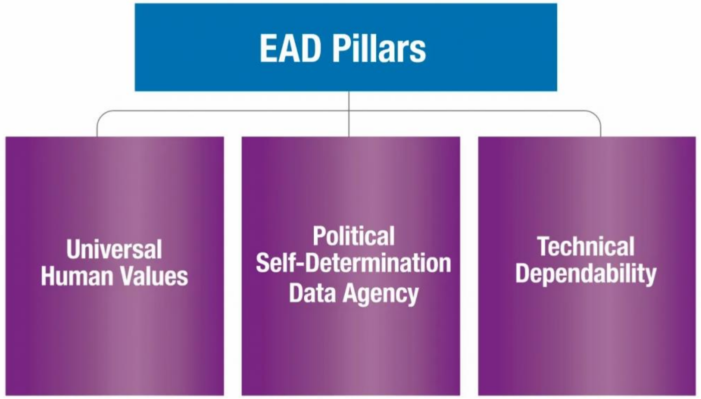
ETHICS
IN
AI

The 3 Pillars of Ethically Aligned Design

- **Universal Human Values:** Design guidelines to respect human rights, protect the environment and align with human values, and increase overall human well-being while empowering as many people as possible.

- **Political Self-Determination and Data Agency:** If people have control over data relating to their personal identity, it will help nurture political freedom and democracy, in accordance with the cultural precepts and values of individual societies.

- **Technical Dependability:** People should be able to trust the services they receive from AIS. They should believe that the system will do its intended function without bias and while respecting human values. Technologies should be audited and monitored to ensure that their operation meets predetermined ethical objectives aligning with human values and respecting codified rights.



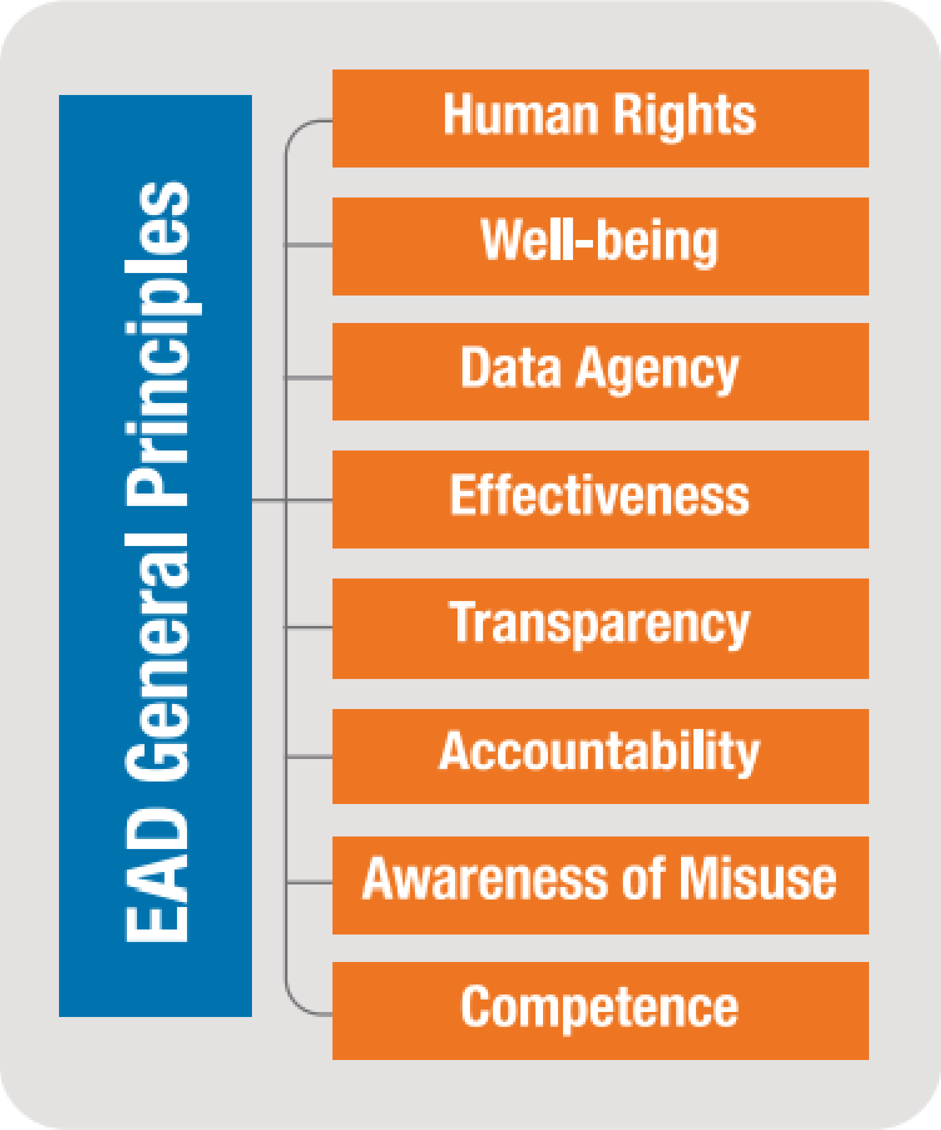
EAD Pillars

Universal Human Values

Political Self-Determination Data Agency

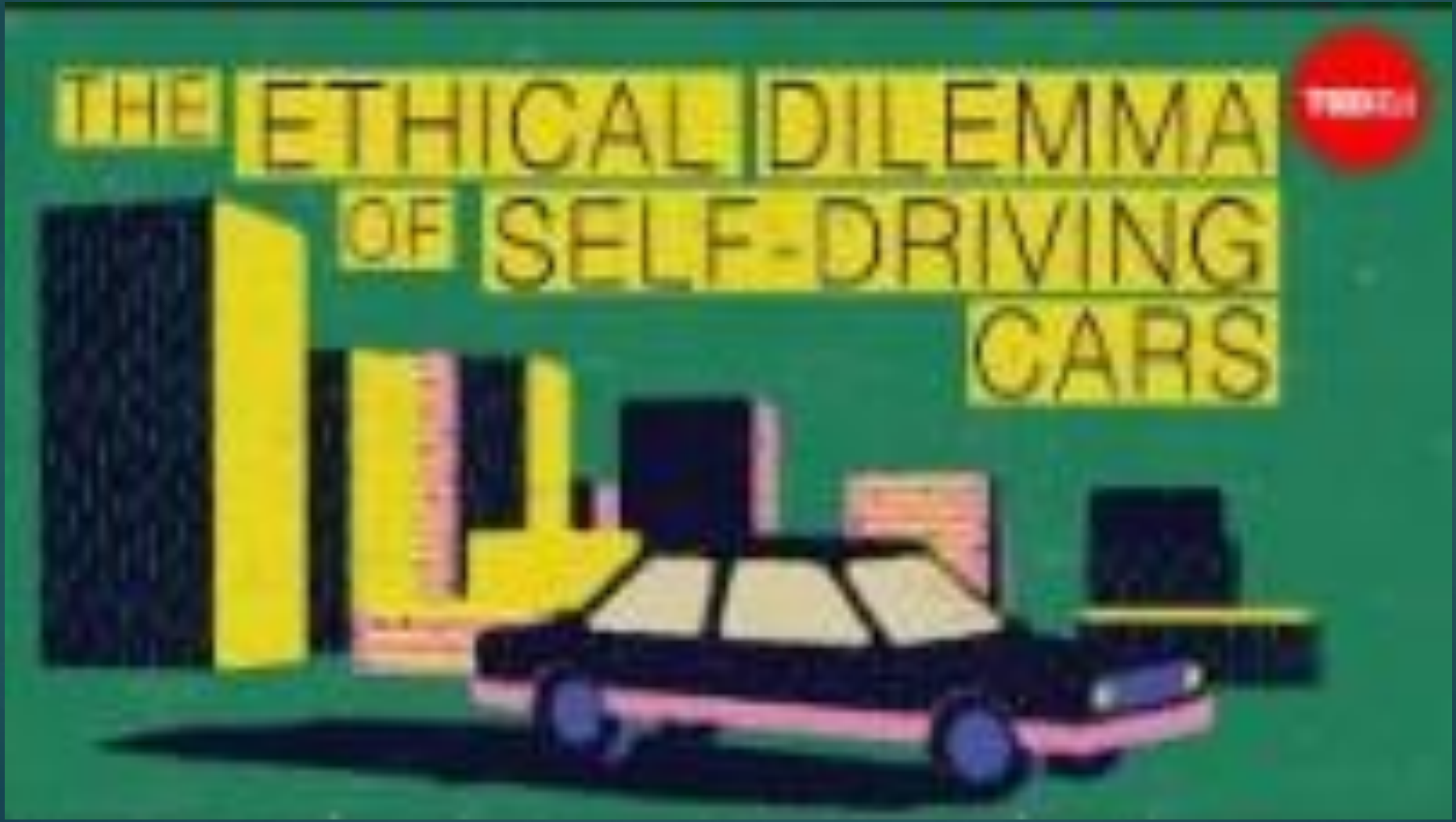
Technical Dependability

The General Principles of Ethically Aligned Design



EAD General Principles

- Human Rights
- Well-being
- Data Agency
- Effectiveness
- Transparency
- Accountability
- Awareness of Misuse
- Competence


THE ETHICAL DILEMMA
OF SELF-DRIVING
CARS

The General Principles of Ethically Aligned Design

1. **Human Rights**–A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being–A/IS** creators shall adopt increased human well-being as a primary success criterion for development.
3. **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.
4. **Effectiveness**–A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency**–The basis of a particular A/IS decision should always be discoverable.
6. **Accountability**–A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse**–A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence**–A/IS creators shall specify, and operators shall adhere to the knowledge and skill required for safe and effective operation.

Purpose behind these General Principles of Ethically Aligned Design

- Embody the highest ideals of human beneficence within human rights.
- Prioritize benefits to humanity and the natural environment from the use of A/IS over commercial and other considerations. Benefits to humanity and the natural environment should not be at odds—the former depends on the latter. Prioritizing human well-being does not mean degrading the environment.
- Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems, in particular by ensuring actions of A/IS are accountable and transparent.

Mapping the Pillars to the Principles

		EAD Pillars		
		Universal Human Values	Political Self-Determination Data Agency	Technical Dependability
EAD General Principles	Human Rights	■	■	
	Well-being	■	■	
	Data Agency	■	■	■
	Effectiveness			■
	Transparency	■	■	■
	Accountability	■	■	■
	Awareness of Misuse			■
	Competence			■

A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights

- To implement human rights ideals within the business, in turn these ideals will be incorporated when coding AI systems.
- The **United Nations Guiding Principles on Business and Human Rights** (UNGPs) - The first global standard for preventing and addressing the risk of adverse impacts on human rights linked to business activity, and continue to provide the internationally accepted framework for enhancing standards and practice regarding business and human rights.

UNITED NATIONS
GUIDING PRINCIPLES
ON BUSINESS
& HUMAN RIGHTS

PRINCIPLE 1: HUMAN RIGHTS - Recommendations

- Further developing governance frameworks will improve public trust in AI/S
- A way to translate current & future legal obligations into technical considerations which can be implemented into AI development
- 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

One would argue that most of the most common and popularly used AI

PRINCIPLE 2 : Well-being

A/IS creators shall adopt increased human well-being as a primary success criterion for development.

- How can AI/S demonstrate an improving of benefit to human society without it just being about economic benefit or lack of negative consequences?
- Well-being, for the purpose of Ethically Aligned Design, is based on the Organization for Economic Co-operation and Development's (OECD) "Guidelines on Measuring Subjective Well-being" perspective that, "Being able to measure people's quality of life is fundamental when assessing the progress of societies."
- AI/S technologies created with the best of intentions can have negative impact on human well being if taken out of context and used for purposes other than its original intended purpose. E.g. People tracking used for healthcare settings can also be used to track people by the military or intelligence services

PRINCIPLE 2 : 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

PRINCIPLE 3 - 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.

- even when individuals provide consent, the understanding of the value regarding their data and its safety is out of an individual's control. E.g. Cambridge Analytica Scandal
- People don't know how their data is being used at all times or when predictive messaging is honoring their existing preferences or manipulating them to create new behaviors.
- Regulations like the EU General Data Protection Regulation (GDPR) will help, but until individuals have some way to better control the use of their personal data.

PRINCIPLE 3 - Data Agency **Recommendations**

- 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. How this can be implemented in practice is the challenge
- The thinking is that there should be technology to give the individual 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.

PRINCIPLE 4 : Effectiveness Creators and operators shall provide evidence of the effectiveness and fitness for purpose of AI/S.

- A/IS will not be trusted unless they can be shown to be effective in use. Any harm to a human will undermine progress made towards convincing the public of the Ethicality of an AI/S
- A meaningful, actionable, accurate and effective method of measuring the effectiveness of an AI/S is important
- Guidance of how to interpret the effectiveness criteria is equally important for the evidence to be credible

PRINCIPLE 4 : Recommendations

- Define clearly benchmarks to measure effectiveness of the system in meeting its objectives, adhering to standards and remaining within risk tolerances.
- Guidance on Interpreting results of these testing
- 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.
- 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

Principle 5 :Transparency

The basis of a particular A/IS decision should always be discoverable.

- How and why a system made a particular decision
- Must be transparent to a wide range of stakeholders for different reasons :
  - 1. For users, what the system is doing and why.
  - 2. For creators, including those undertaking the validation and certification of A/IS, the systems’ processes and input data.
  - 3. For an accident investigator, if accidents occur.
  - 4. For those in the legal process, to inform evidence and decision-making.
  - 5. For the public, to build confidence in the technology

Principle 5 : Transparency **Recommendations**

- New standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined. Again, for different stakeholders,
  - 1. 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.
  - 2. For validation or certification agencies, the algorithms underlying the A/IS and how they have been verified.
  - 3. For accident investigators, secure storage of sensor and internal state data comparable to a flight data recorder or black box.
- IEEE P7001™, IEEE Standard for Transparency of Autonomous Systems is one such standard

The EU's new Regulation on AI
Europe Calls for More Transparency
F

PRINCIPLE 6 : Accountability

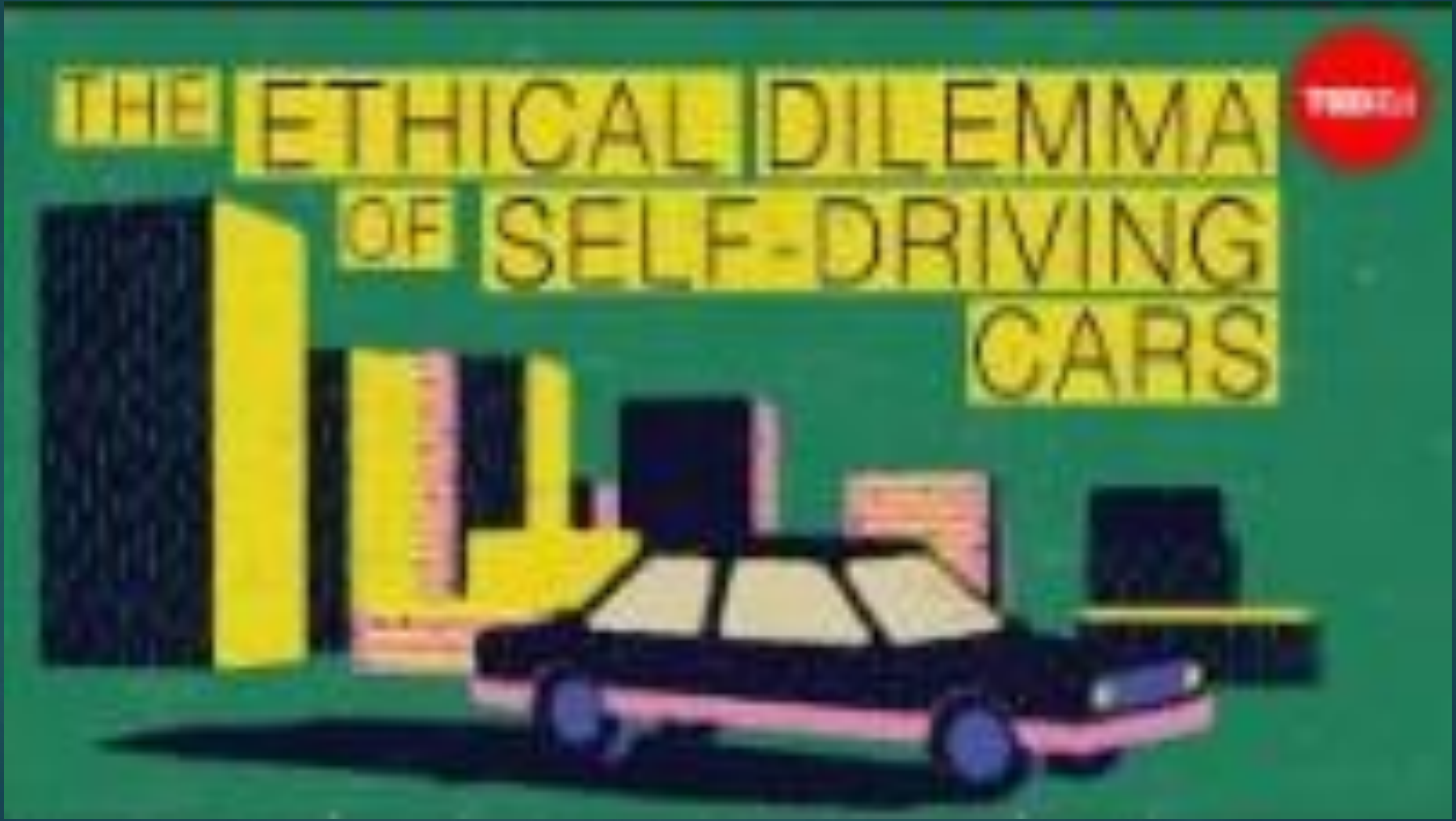
A/IS shall be created and operated to provide an unambiguous rationale for decisions made.

- Manufacturers of these systems must be accountable in order to address legal issues of culpability.
- Accountability and partial accountability are not possible without transparency, thus this principle is closely linked with Principle 5

Recommendations

- Legislatures/courts should clarify responsibility, culpability, liability, and accountability for A/IS, where possible, prior to development and deployment
- Designers and developers of A/IS should remain aware of existing cultural norms
- As this is a new fiels, where there are no applicable norms, Multi-stakeholder ecosystems including creators, and government, civil, and commercial stakeholders should be formed to develop such norms
- 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.

Accountability ?



Principle 7 :Awareness of Misuse

Creators shall guard against all potential misuses and risks of A/IS in operation.

- New technologies give rise to greater risk of deliberate or accidental misuse
- Cases of A/IS hacking have already been widely reported, The Microsoft Tay AI chatbot was famously manipulated when it mimicked deliberately offensive users
- Responsible innovation requires A/IS creators to anticipate, reflect, and engage with users of A/IS.

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 by:
  - Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS.
  - Having credible experts educate the public to reduce unwarranted fear of AI/S
  - Educate govt/lawmakers/regulatory agencies about AI/S so that it will reassure citizens that their interests are protected.

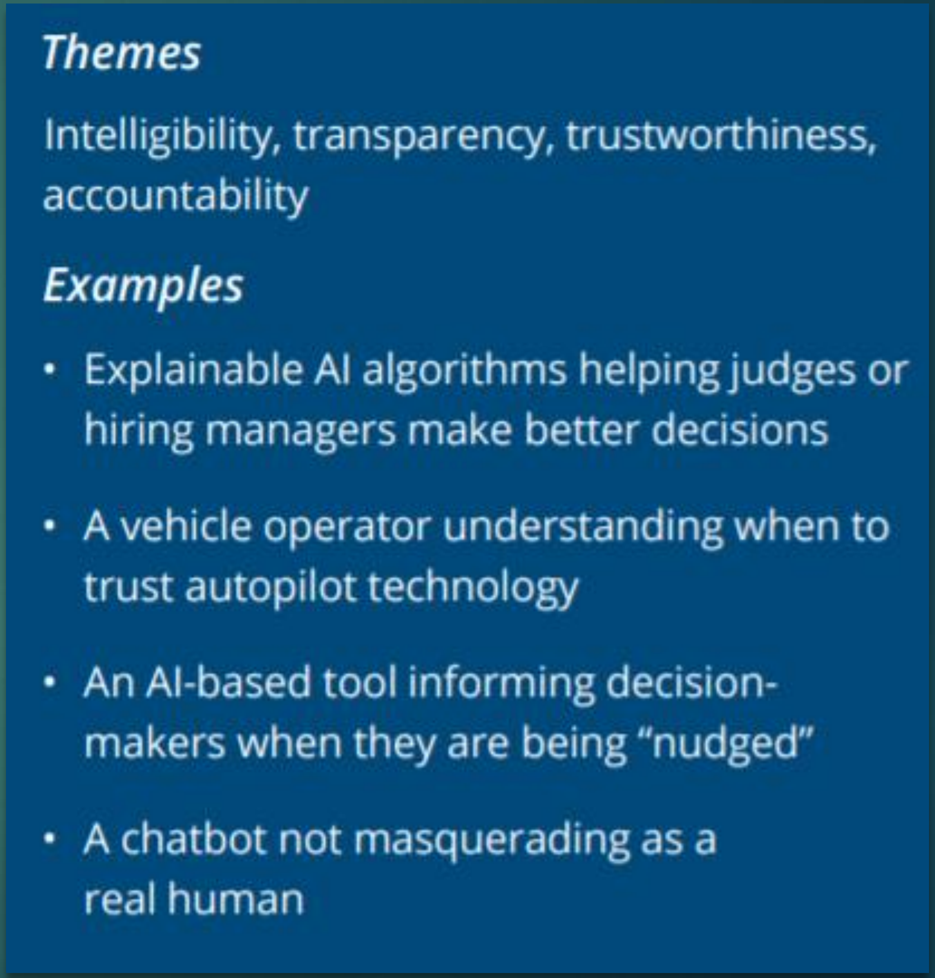
Principle 8 :Competence

Creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

- As decisions become more complex, Operators of A/IS can become less likely to question and potentially less able to question the decisions that algorithms make.
- Operators should be able to understand how A/IS reach their decisions, the information and logic on which the A/IS rely, and the effects of those decisions. Even more crucially, operators should know when they need to question A/IS and when they need to overrule them. Creators of AI/S should take special care to consider and facilitate these into their systems

Principle 8 : Competence **Recommendations**

- 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.
- 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.
- Documented policies to govern how A/IS should be operated should be available for operators.
- Operators of A/IS should, before operating a system, make sure that they have access to the requisite competencies.



Themes

Intelligibility, transparency, trustworthiness, accountability

Examples

- Explainable AI algorithms helping judges or hiring managers make better decisions
- A vehicle operator understanding when to trust autopilot technology
- An AI-based tool informing decision-makers when they are being "nudged"
- A chatbot not masquerading as a real human

Classical Ethics in AI/S

- AI/S are autonomous, amoral systems. What are the implications of this fact on the ethics of decisions made by such systems? Decisions which have moral consequences.
- The goal is to ensure that decisions reached by AI/S mimic the decisions made by humans
- In an attempt to do this, it is suggested that classical ethics methodologies (drawing from phylosophical, religious & cultural)are used to try and guide the morality of the decisions an AI/S makes
- responsibility and accountability for the decisions made by autonomous systems is what is being attempted.

Researchers at the University of Cambridge’s Leverhulme Centre for the Future of Intelligence (LCFI) have been awarded nearly two million Euros to build a better understanding of how AI can undermine “core human values”.

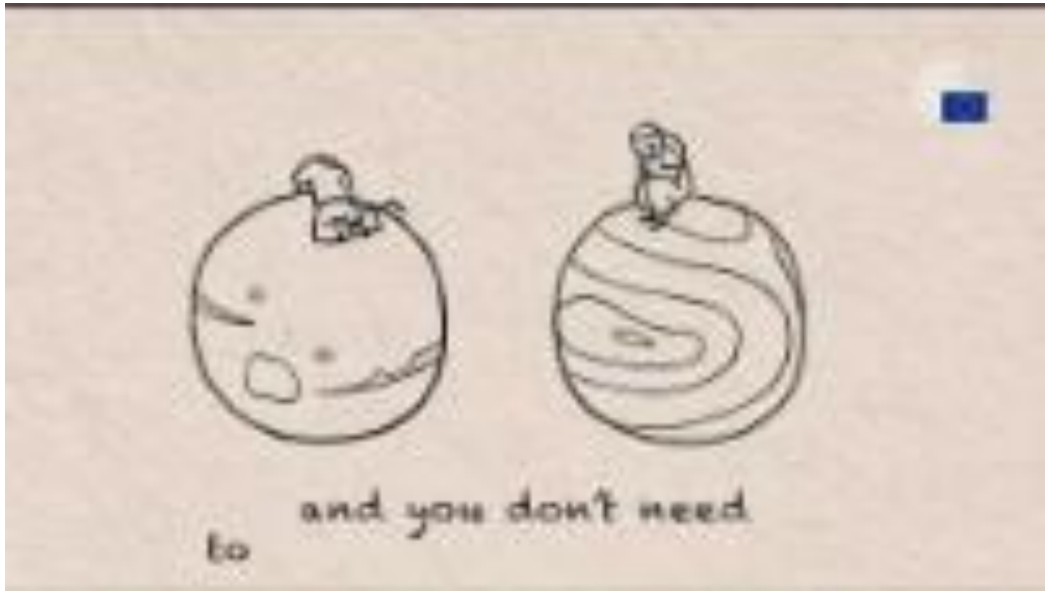
The grant will allow LCFI and its partners to work with the AI industry to develop anti-discriminatory design principles that put ethics at the heart of technological progress.

assess, understand, measure, monitor, safeguard, and improve the **WELL-BEING** impacts of A/IS on humans

- Today, A/IS creators largely measure success using metrics including profit, gross domestic product (GDP), consumption levels, and occupational safety.

- Consistent and multidimensional indicators are needed that are easily implementable by the developers, engineers, and designers who are building our future.

- Example: Self driving cars also have the potential to help the environment by reducing greenhouse gas emissions and increasing green space. Autonomous vehicles can also positively impact wellbeing by increasing work-life balance and enhancing the quality of time spent during commutes.



Affective Computing (Emotion AI)

- **Affective computing** is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. In general, affective computing refers to a computer's capabilities **to recognize a user's emotional states, to express its own emotions, and to respond to the user's emotions** (Picard, 1997).
- A/IS should be used to help humanity to the greatest extent possible in as many contexts as are appropriate. However, there is also potential that harm could be caused either by amplifying, altering, or even dampening human emotional experience.
- **Example:** Automotive companies can leverage computer vision to track the driver's emotional state while driving. If the driver is too tired, stressed, or angry/sad, it can provide alerts for unsafe driving. However, ***The Ministry of Happiness*** in the United Arabic Emirates has started an initiative to understand the general mood of the population using video analysis cameras in public places. For some, this is crossing the line where something very personal such as emotion is recorded without permission.

Personal Data and Individual Agency

- Humans cannot respond on an individual basis to every algorithm tracking their behavior without technological tools supported by policy allowing them to do so.
- Individuals may provide consent without fully understanding specific terms and conditions agreements.
- they are also not equipped to fully recognize how the subtle use of their data to inform personalized algorithms affects their choices at the risk of eroding their ability to make an informed decision on how it could influence and shape their life trajectory.
- To strengthen individual agency, governments and organizations must test and implement technologies and policies that let individuals create, adjust and control their online agency as associated with their personal identity.

Methods to Guide Ethical Research and Design

- To ensure that A/IS benefit humanity, A/IS research and design must be underpinned by ethical and legal norms. These should be done through values-based research and design methods which put human well-being at the core of A/IS development.
- To help achieve these goals, researchers, product developers, and technologists across all sectors need to embrace research and development methods that evaluate their processes, products, values, and design practices in light of the concerns and considerations raised by various organisations outlining the potential harm which can be caused by AI/S

A/IS for Sustainable Development

- The scaling and use of A/IS represent a genuine opportunity across the globe to provide individuals and communities, rich, poor, rural or urban with the means to satisfy their needs and develop their full potential,
- A/IS will potentially disrupt economic, social, and political relationships and interactions at many levels. Those disruptions could provide an historical opportunity to reset those relationships in order to distribute power and wealth more equitably and thus promote social justice.
- The ethical imperative driving this section is that A/IS must be harnessed to benefit humanity, promote equality, and realize the world community’s vision of a sustainable future and the SDGs (The 17 Sustainable Development Goals as set by the UN with a target date of 2030)

Embedding Values into Autonomous and Intelligent Systems

- There are no universal rules on embedding human values and norms into A/IS. Yet, such systems are instilled with increasing autonomy in making decisions and manipulating their environment. Therefore, it is essential that they are designed to adopt, learn, and follow the norms and values of the community they serve.
- A community's network of social and moral norms is likely to reflect the community's values, and A/IS equipped with such a network would, therefore, also reflect the community's values. To do this with AI/S,
  1. Identifying the norms of the specific community in which the A/IS operate
  2. Computationally implementing the norms of that community within the A/IS, and
  3. Evaluating whether the implementation of the identified norms in the A/IS are indeed conforming to the norms reflective of that community.

Policy

- To encourage the development of socially beneficial applications of A/IS, and to protect the public from adverse consequences of A/IS, intended or otherwise, effective policies and government regulations are needed.
- Such policies protect and promote safety, privacy, human rights, and cybersecurity, as well as enhance the public's understanding of the potential impacts of A/IS on society.
- effective A/IS policies should embody a rights-based approach1 that addresses five issues:
  - Ensure that A/IS support, promote, and enable internationally recognized legal norms.
  - Develop government expertise in A/IS.
  - Ensure governance and ethics are core components in A/IS research, development, acquisition, and use.
  - Create policies for A/IS to ensure public safety and responsible A/IS design.
  - Educate the public on the ethical and societal impacts of A/IS.

Law

- When AI/S is used in everyday life, it is affected by the Law. It also affects the Law. Science, technological development, law, public policy, and ethics are tied to each other and is what creates social order.
- The Law not only responds to the technological innovation represented by A/IS, but also on how the law guides and sets the conditions for that innovation.
- With this complex interactive relationship between the Law and AI/s. we seek to identify principles that will steer the process in a manner that leads to the improvement, prosperity, and well-being of everyone

When AI
Gets Too Smart

The Mission and Results of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

- To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity