The State of Ethical AI Frameworks

Making sense of worldwide initiatives in Ethical AI

Presented by:

Kathleen Walch
Ronald Schmelzer
Megan Yarbrough

About Cognilytica

- Cognilytica is an ***AI-focused analyst and advisory firm***
- We are a small business, based in Maryland
- Market research, advisory & guidance on ***AI, ML, & Cognitive Technology***
- Provides ***role-specific education*** on AI, ML, and emerging technology
- Focused on ***enterprise and public sector adoption*** of AI, ML, and Cognitive Technology
- Kathleen Walch and Ron Schmelzer are ***Principal Analysts and Managing Partners*** and Megan Yarbrough is a ***Research Associate*** at Cognilytica.
- Produce the popular ***AI Today podcast*** and ***AI communities including AI in Government and Data for AI***

Some Caveats

- We are not ethicists nor experts in technology ethics
- Definitions are imprecise and often conflicting
- No widespread consensus for terminology or definitions
- Boundaries between concepts are fuzzy
- Many concepts are highly culture and context specific

… however, organizations have many of the same challenges, and despite those issues, they are making decisions and creating frameworks…

Tackling the Fears and Concerns of AI (And Big Data)

AI Fears	AI Concerns
• Intelligent systems will take over the world • Robots will take my job • I'll lose control over my privacy and data • Too much data in too few hands • I'll live in a surveillance state • The robots will kill me	• Lack of transparency in algorithmic decision-making • Bad actors doing bad things with AI • AI systems vulnerable to tampering and data security issues • Systems are susceptible to bias • Laws not keeping up with technology • How can I really trust these systems?

Ethical AI: “Right vs. Wrong”

AI Ethics: “AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies” [1]

Bad Machines	Bad People Doing Bad Things	Bad Practices	Bad Visibility
• Threatening lives • Threatening freedoms • Threatening control • Threatening dignity • Threatening environment • Not in best interests of humanity	• Violating our laws • Violating our trust • Violating our privacy • Violating our lives	• Lack of Safety • Lack of accountability • Lack of positive purpose • Lack of care for workforce disruption	• Limited visibility into data and processes • Limited disclosure • Limited consent • Limited visibility into algorithmic behavior • Limited repeatable, consistent processes

Some Inspiration from the Past (1942)

First Law

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.

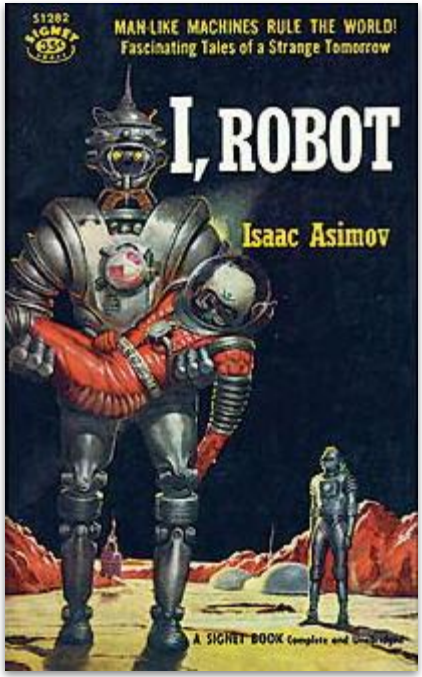
Second Law

- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

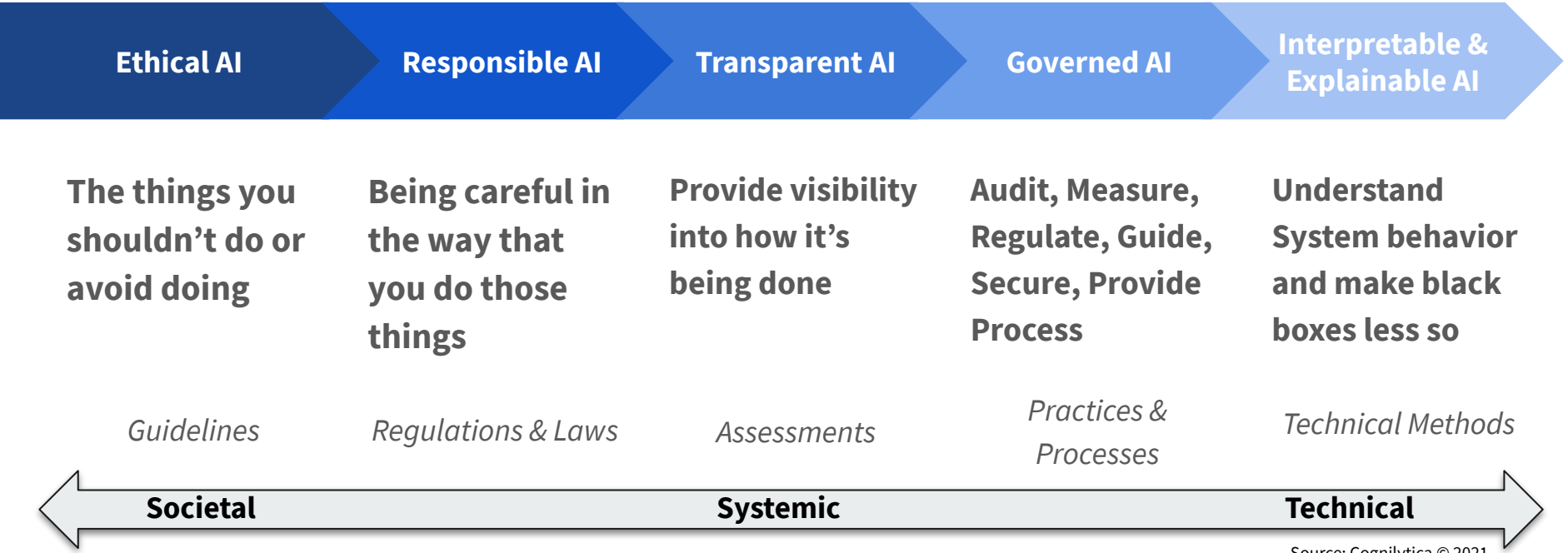
Third Law

- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

But this is clearly not enough. And it's also science fiction.



Ethical AI Concepts: A Spectrum

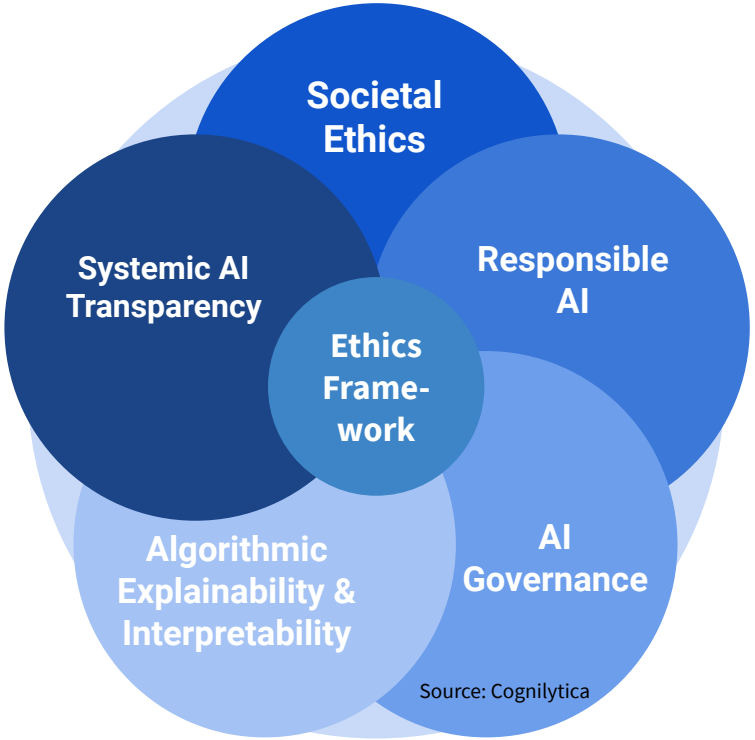


Source: Cognilytica © 2021

Ethical AI Frameworks: A Bit Of Everything

- Cognilytica analyzed over **60 National, Organizational, and Corporate Ethical AI Frameworks**
- Mostly a ***mish-mash of ethical concepts***
- The words they use often ***don't match the ethical concept***
- There are ***no comprehensive ethical AI frameworks***
- The ***gaps in ethical AI frameworks*** are eye opening
- ***Mix of recommendations*** for technology users and technology creators

We analyzed them all, normalized the terminology and compared them against each other.



Societal Ethics: The Main Elements

Human Values
- Machine-based systems should exhibit the same values that we have as humans. Do no harm.

Dignity
- AI systems should not treat humans as machines.

Fairness
- AI systems should not favor one group over another

Diversity & Inclusion
- AI systems should be built for and incorporate data from the breadth of humanity

Bias & Discrimination
- AI systems shouldn't further bias or discrimination

Freedom & Agency
- AI systems shouldn't limit human choice or freedom of action.

Human Benefit
- AI systems should be built for the benefit of the widest group of humanity, and not for the benefit of a few.

Human Control
- AI systems should never operate without humans in control

Respect of the Environment
- AI systems should take care not to abuse the environment. Do no harm.

Societal Ethics: How the Ethics Frameworks Stack up

Canada:
"Montréal Declaration for Responsible Development of Artificial Intelligence"
Very complete, combine with 2 other Canadian efforts

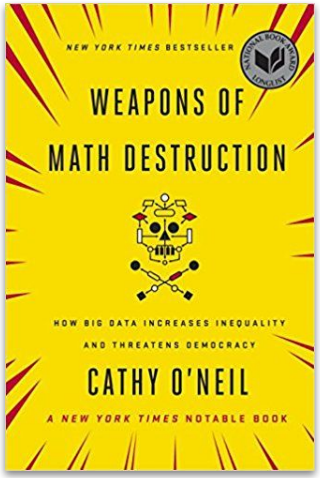
Microsoft:
"Microsoft Responsible AI Principles"
Surprisingly a very strong ethical framework.
Not only provides principles, but methods to ensure their application

China:
Ministry of Science & Technology:
"Development of responsible artificial intelligence"
Unusual terminology
"Harmony & Friendship; Share responsibilities; Tolerance and sharing; Fairness and justice"

USA DoD JAIC:
"Ethical Principles for Artificial Intelligence"
Surprisingly weak in societal ethics.
"Equitable (Bias only); Governable (Human Control)"

What is Responsible AI?

- **Responsible AI:**
  - "AI solutions that help maintain individual trust and minimize privacy invasion. Responsible AI places human (e.g., end-users) at the center"[2]
- Just because you can do it, even ethically, how can you do it the ***right*** way?
- How can AI be done in a way that doesn't violate people's trust?
- How can misusing AI technology be avoided?
  - ***"Don't misuse or abuse AI technology"***
- Examples of Responsible AI Concerns
  - Facial recognition as responsible vs. irresponsible AI
  - Responsible vs. Irresponsible Algorithmic decision-making
  - Development of invasive user profiles
- **What is a Responsible AI Framework?**
  - Set guidelines for the proper use of AI and guardrails to prevent abuse or misuse



Responsible AI: The Main Elements

★ Positive Purpose

- AI systems must be built for some positive purpose

★ Safety & Security

- AI systems should be safe and secure

★ Trust

- AI systems should not violate human trust or cause people to mistrust entities

★ Human Accountability

- Human individuals should be identified who are responsible and accountable for the behavior and operation of the AI systems

★ Privacy

- AI systems should not violate the privacy of humans or impose state-wide surveillance

★ Misuse, Abuse, and Compliance with Laws

- Humans should not misuse AI systems for any criminal or non-law abiding purpose, or to use AI to circumvent laws or regulations.

★ Lethal Autonomous Weapons

- AI systems should not be built to enable lethal autonomous weapons

★ Workforce Disruption

- AI systems should not be built that have as an intentional goal the mass replacement of human workers, disruption to economies

Responsible AI: How the Ethics Frameworks Stack up

AI4People:
"Ethical Framework for a Good AI Society"
The most complete with respect to Responsible AI principles

Workforce Disruption:
Only a few frameworks talk about avoiding workforce disruption as an ethical goal.
China; India; OECD; European and African.

Multinational and Org Efforts beat Country Efforts:
Countries are not paying enough attention to Responsible AI uses. Especially around AI applied to weapons.

NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon (FAI):
Details needed on Responsible AI
Not really an ethical AI framework, but it could be...

What is AI System Transparency?

- **AI Transparency**:
  - ***Visibility into all the aspects of what went into building an AI system so users can understand the full context of how an AI system is built and used***
- ***How do I know what went into making this AI system so I can trust it?***
  - How do I know how the AI models are being built, what data has been used to create the models, what data was being included / excluded, and aspects of bias?
- Conflicting definitions of transparency
  - ***Transparency as Interpretability / Explainability? "Algorithmic Transparency"***
    - That's algorithmic transparency, which has to do with how interpretable or explainable the algorithm is
    - Detailed separately.
  - ***Transparency as Visibility into how the system was built? "Systemic Transparency"***
    - It's even more important to know what data went into making that model in the first place.
    - ***Especially if it's not working how you think it should!***
- ***"The ImageNet test set has an estimated label error rate of 5.8%"*** [5]
  - A mushroom is labeled a spoon, a frog is labeled a cat, a high note from Ariana Grande is labeled a whistle.
  - Who cares if the model explains why it said "turn right" if the map was wrong
  - ***Without transparency, no visibility!***

Error-riddled data sets are warping our sense of how good AI really is

AI Systemic Transparency: The Main Elements

★ AI System Transparency

- AI systems should provide visibility into the data and components of the system with their configuration that is used to generate results
- Human decisions on the operation, versioning, development, and use of the AI system should be disclosed and open.

★ Bias Measurement & Mitigation

- AI systems should provide a means to constantly measure bias of various sources and provide means to mitigate any bias detected.

★ Open Systems

- AI systems as a whole should use open source technology with the mechanism by which the system operates visible to all

★ Disclosure & Consent

- Organizations should disclose when AI systems are being used and when humans are interacting with AI systems
- AI systems should provide a means for humans to consent out of interaction with AI systems, being included in AI models, or otherwise being impacted by the AI system.

Systemic Transparency: How the Ethics Frameworks Stack up

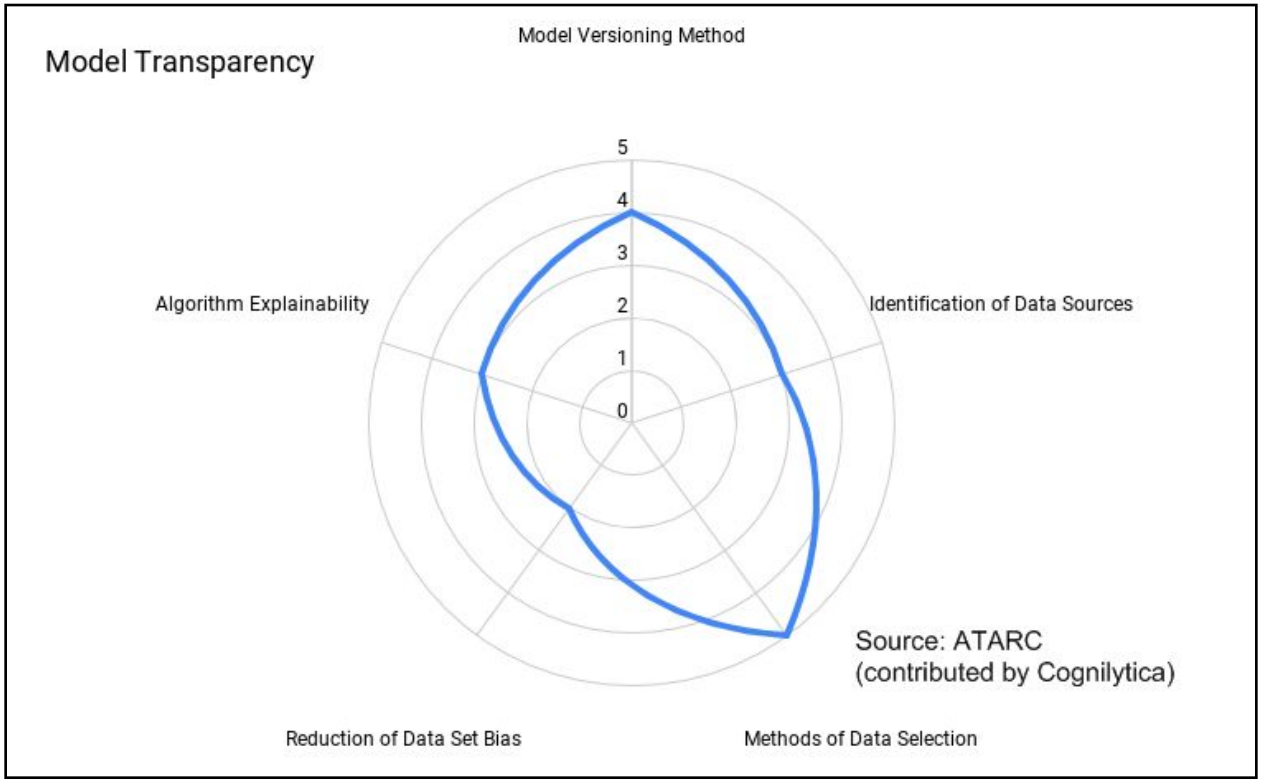
AI System Transparency:
Many ethics frameworks talk about total AI system visibility and transparency.
Need a way to measure.

Corporate Ethical AI frameworks mostly missing systemic AI transparency:
Missing disclosure & consent, visibility into system operation, measurement of bias, and open systems

Academic ethical AI frameworks most concerned about openness:
Frameworks originating within university and research settings most often mention openness in research, data, shared platforms, and systemic AI transparency

USA DoD JAIC:
"Ethical Principles for Artificial Intelligence"
Completely absent of any systemic transparency principles

Assessing AI Transparency: Model Assessment Approach



AI Governance: The Main Elements

System Auditability

- AI systems should provide ways to audit all aspects of operation and behavior

Contestability

- AI systems need to provide ways to contest or appeal AI decisions for human review.

Risk Assessment & Mitigation

- Organizations need established methods to assess ongoing risk to AI systems and identified means to mitigate those risks

System Monitoring & Quality

- Systems operating within ethical guidelines need to make sure that AI systems are always operating within acceptable performance, usage, and other parameters

Education & Training

- Ethical frameworks should require those who are involved in AI system creation or use to be trained in their proper development and use.

Regulation & Certification

- AI systems should comply with requirements of regulatory bodies and third-party certifications with regular third-party audits and certifications of ethical operation

AI Governance: How the Ethics Frameworks Stack up

Canada
"Directive on Automated Decision-Making"
Addresses almost all of the points of AI Governance

European Union
Various efforts
Spend the most amount of time talking about regulation, certification, and third-party credentialing or accountability

An Informed Workforce
The topic of education and training of the workforce on AI comes up frequently in ethical AI frameworks

A lot of talk about bias, but little governance effort
Only a few ethical AI frameworks require organizations to measure the extent to which AI systems exhibit bias

What is Interpretable AI and Explainable AI?

- Discomfort with “Black Boxes”
- **Explainable AI:**
  - **Can the model tell you how it arrived at a particular prediction / conclusion?**
  - “An explainable model is a … method/technique to be able to peer into the black-box and understand how the model works”[5]
- **Interpretable AI:**
  - **If not, can you provide ways to understand the components of the decision-making or the key factors contributing to a result?**
  - “Interpretability is the degree to which a human can understand the cause of a decision” or “The degree to which a human can consistently predict the model’s result”[3]
- **“Root Cause Analysis” / “Failure Analysis”**
- Can we understand how algorithmic decisions are being made?
  - **“Algorithmic Transparency”**
- How do I know what the AI system is doing?
- Do I have visibility into how the systems are succeeding or failing and why?
- Is it possible to “debug” AI systems?

AI Explainability & Interpretability: The Main Elements

★ Understandability / Root Cause Explanations

- When AI systems fail to provide expected results, AI systems should always provide a human understandable means to understand the root cause of any failures. Explanations without necessarily algorithmic explanation.

★ Algorithmic Interpretability

- AI systems should provide a means to interpret AI results so that cause and effect can be understood, even with limited algorithmic explainability

★ Algorithmic Explainability

- AI systems should use algorithms that provide a direct means to explain how outcomes were arrived from input data

Algorithmic Explainability: How the Ethics Frameworks Stack up

Understandability vs. Explainability

Acknowledgement of the difficulties of algorithmic explainability

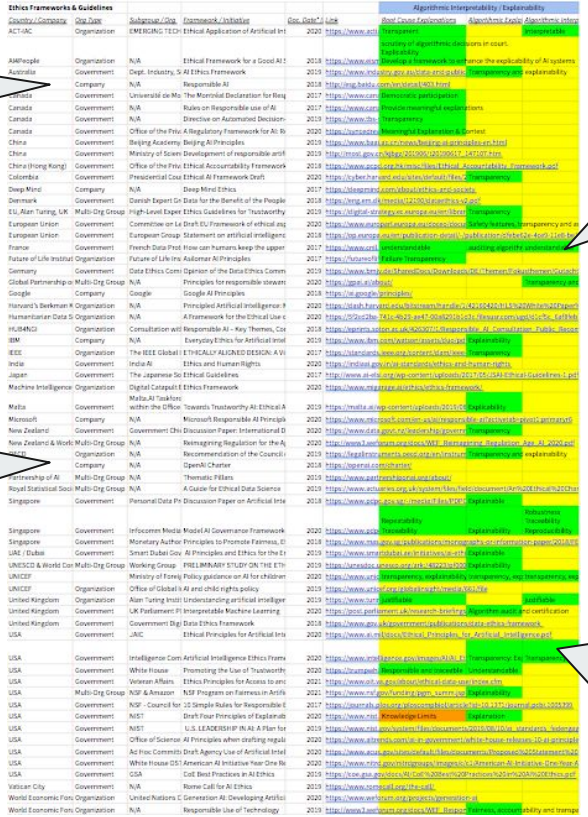
"Common sense explanations"

India

"India AI: Ethics and Human Rights"

Provides a good definition of goals of understandability and explanation

Focus is on understandability for the purposes of giving users disclosure and ability to contest results



France

"French Data Protection Authority (CNIL)"

Checks all the boxes for understandability, interpretability,and explainability

USA (White House)

"Promoting the Use of Trustworthy Artificial Intelligence in Government (EO13960)"

Does a good job of identifying needs for explainability in an ethical AI Framework

"Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate."

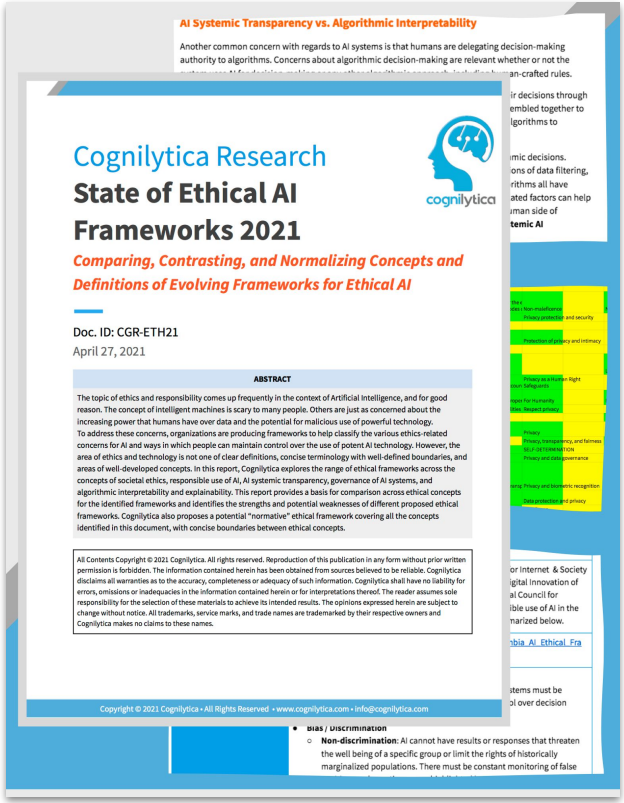
In Search of a Comprehensive Ethical AI Framework

… wait… we just presented it!

Societal Ethics	Responsible AI	Systemic AI Transparency	AI Governance	AI Explainability & Interpretability
★ Human Values ★ AI for Human Benefit ★ Dignity ★ Fairness ★ Respect Diversity ★ Avoid Bias ★ Ensure Freedom & Human Agency ★ Keep the Human in Control ★ Respect the Environment	★ Positive Purpose ★ Law Abiding use of AI ★ Ensure Safety ★ Maintain Trust ★ Provide Human Accountability ★ Ensure Privacy ★ No Lethal Autonomous Weapons ★ Avoid Workforce Disruption	★ Visibility into AI Data ★ Bias Measurement & Mitigation ★ Open Systems & Human Decisions ★ Provide Disclosure and Respect Consent	★ Risk Assessment & Mitigation ★ System Auditability ★ Contestability ★ Continuous System Quality ★ External Regulatory Bodies & Third-party Certifications ★ Training	★ Algorithmic Explainability ★ Algorithmic Interpretability ★ Root Cause Explanations

Cognilytica's State of Ethical AI Frameworks 2021 Report

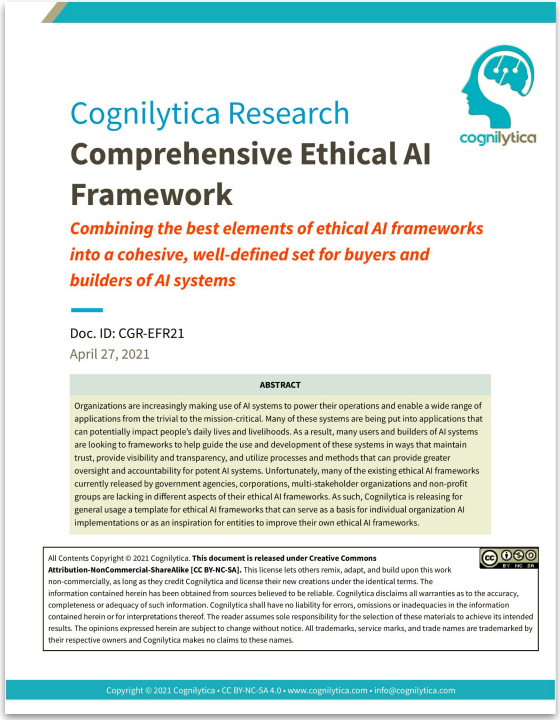
- ***State of Ethical AI Frameworks 2021*** report published April 2021
- Thorough evaluation of over ***60 ethical AI Frameworks***
- Provides ***normalized definitions and means for comparison***
- ***Guided Questions*** for buyers and builders of AI systems
- This report is 150+ pages with comprehensive overviews of each AI framework. It's available for Cognilytica research subscribers to download or for individual purchase.



Cognilytica Comprehensive Ethical AI Framework

- In addition, we produced ***Cognilytica's Comprehensive Ethical AI Framework*** as a free framework that serves as a guide for others.

Get a copy of Cognilytica's Comprehensive Ethical AI Framework



Cognilytica Research

Comprehensive Ethical AI Framework

Combining the best elements of ethical AI frameworks into a cohesive, well-defined set for buyers and builders of AI systems

Doc. ID: CGR-EFR21

April 27, 2021

ABSTRACT

Organizations are increasingly making use of AI systems to power their operations and enable a wide range of applications from the trivial to the mission-critical. Many of these systems are being put into applications that can potentially impact people's daily lives and livelihoods. As a result, many users and builders of AI systems are looking to frameworks to help guide the use and development of these systems in ways that maintain trust, provide visibility and transparency, and utilize processes and methods that can provide greater oversight and accountability for potent AI systems. Unfortunately, many of the existing ethical AI frameworks currently released by government agencies, corporations, multi-stakeholder organizations and non-profit groups are lacking in different aspects of their ethical AI frameworks. As such, Cognilytica is releasing for general usage a template for ethical AI frameworks that can serve as a basis for individual organization AI implementations or as an inspiration for entities to improve their own ethical AI frameworks.

Thank You

Presenters:

Kathleen Walch, Ronald Schmelzer, Megan Yarbrough

Cognilytica - http://www.cognilytica.com

info@cognilytica.com

Appendix / Sources

Sources

1. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529
2. Wang, Yichuan & Xiong, Mengran & Olya, Hossein. (2019). Toward an Understanding of Responsible Artificial Intelligence Practices. 10.24251/HICSS.2020.610.
3. Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics. 2019; 8(8):832. https://doi.org/10.3390/electronics8080832
4. https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48
5. “Error-riddled data sets are warping our sense of how good AI really is” - https://www.technologyreview.com/2021/04/01/1021619/ai-data-errors-warp-machine-learning-progress/