

# **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)





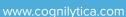
Al Fears	AI Concerns
<ul> <li>Intelligent systems will take over the world</li> <li>Robots will take my job</li> <li>I'll lose control over my privacy and data</li> <li>Too much data in too few hands</li> <li>I'll live in a surveillance state</li> <li>The robots will kill me</li> </ul>	<ul> <li>Lack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



# 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"<sup>1</sup>

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





# 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.

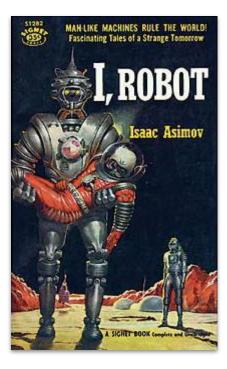
#### Second Law

• 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**

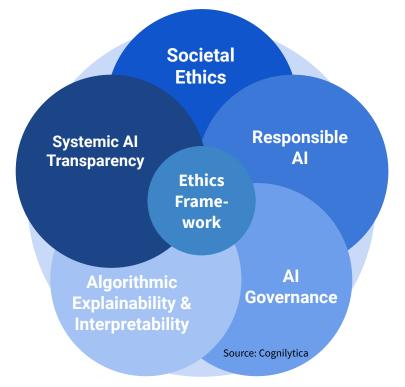
Ethical AI	Responsible AI	Transparent Al	Governed Al	Interpretable & Explainable AI
The things you shouldn't do or avoid doing	Being careful in the way that you do those things	Provide visibility into how it's being done	Audit, Measure, Regulate, Guide, Secure, Provide Process	Understand System behavior and make black boxes less so
Guidelines	Regulations & Laws	Assessments	Practices & Processes	Technical Methods
Societal		Systemic		Technical
N				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

• Al systems shouldn't further bias or discrimination

#### **Freedom & Agency**

AI systems shouldn't limit human choice or freedom of action.

## Human Benefit

Al 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**

 Al 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)"

#### www.cognilytica.co

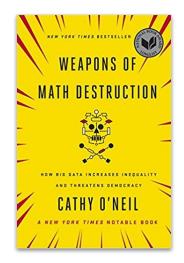




# What is Responsible AI?

- Responsible AI:
  - "AI solutions that help maintain individual trust and minimize privacy invasion. Responsible AI places human (e.g., end-users) at the center"<sup>2</sup>
- 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

	ALERT
Algorithr	ns have consequences.
	ОК



#### 💟 @Cognilytica



# **Responsible AI: The Main Elements**

#### 🖈 🛛 Positive Purpose

Al systems must be built for some positive purpose

## ★ Safety & Security

Al systems should be safe and secure

## \star 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

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

#### У @Cognilytica



# **Responsible AI: How the Ethics Frameworks Stack up**

**Al4People:** *"Ethical Framework for a Good Al Society"* 

The most complete with respect to Responsible AI principles

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

Dies fromeworks & Culdellows Depty Detains On last Internation for Automatic Institution An other Designed States and All Beiling Stationsy Reiling & Principal Lotinger Dougs Their section and high interferen-Alt-american's for the Divisal Line ALCONDAL & TrendsPlan A Curies for 1811 and Data Solaria Storogen Same or Adding into -pairers' 242.12.14 delast Constant inser Alfanta - Bitting Armedia and Ber Annens In and South States Print States of Devilational As the Committe Orall Agency due of without this

Workforce Disruption: Only a few frameworks talk about avoiding workforce disruption as an ethical goal.

China; India; OECD; European and African.

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?

- Al 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%"<sup>5</sup>* 
  - 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 Al really is



#### 💆 @Cognilytica



# 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

#### **r** 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

Al 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

Ethics Frameworks						
Country/Conserv	<u>Pre Lipr</u>	Subgroup / Grg	(remework/initiation	Doc. Detr* (	6lo8	A Sport Dataparts
ACT-IAC	Organization	EMERGING TECH	Ethical Application of Artificial Int	2020	https://www.acti	deteconicable
Tole	Organization	N/8	Ethical Framework for a Good A 5			d autor content con
Autralia .	Covennment	Dept. Industry; S	Al Ethics Framework	2019	https://www.indu	anny good and states and o
Deidu .	Company	N/A	Responsible Al	2015	bttp://ing.baldu.s	amain ideal (MSS Arm
Canada	Coventment	Université de Mo	The Monthial Declaration for Reg	2017	https://www.cats	wan.combheimerten
Canada	Covernment	NA	Rules on Responsible use of Al	2017	https://www.cat	De transparent:
Canada Canada	Coversment Coversment	N/A Office of the Priv	Directive on Automated Decision A Regulatory Framework for Al: Re			tel generalized decranges Generet (2020/12/13/2
China	Coventment		Beijing Al Principles		https://www.bas	
China	Coventment Coventment		Development of responsible artifl Dhical Accountability Framework		http://mest.gov.c	
China (Hang Kong)						dependence Heriter Transportery
Colombia	Coventment		Othical Al Framework Draft		https://cyber.har	
Deep Mind	Company	NA	Deep Mind Ethics	2017	https://deeptaine	Loom Albert Set Inc.
Denmark	Covennment		Data for the Benefit of the People			errodult2190/2atort
EU, Alan Turing, UK	Multi-Org Group		Dhics Guidelines for Trustworthy		https://digital-str	
European Union	Coventment		Draft EU Framework of ethical asp		https://www.eure	
European Union	Covennment		Statement on artificial Intelligence			sester publication-det
France	Coventment		How can humans keep the upper			males dels inflation
Future of Life Institut			Aslornar Al Principles			e orgia principke!
Germany	Coventment		Opinion of the Data Ethics Comm			v.de/SharedDors/Down
Global Partnership o			Principles for responsible stewart		https://gpai.al/ab	
Google	Company	Googie	Google Al Principles	2018	https://sigoode/	arinsisi <i>ni</i> ,
Harvard's Berkman B		NA	Principled Artificial Intelligence: N		betwee lites have	
Humanitarian Data S	Organization	NA	A Framework for the Othical Use of		https://Stictling	7410-4508-9617-00-00
HUDANCI			Responsible Al – Key Themes, Co Transition Drives for Art Exception		https://egrinits.ag	
IEM I	Company Organization	N/A	Everyday Dhics for Art Ecial Intel ETHICALLY ALIGNED DESIGN: A VI		https://www.lbcs	elember de contrate de la contrate de la
ittle	Cryandapon	India Al	ETHICALLY ALIGNED DESIGN: A VI Dhics and Human Rights		https://standards https://indiaal.go	
Japan	Coversment		Ethical Guidelines			
Vapan Hachine Intelligence			Edica Guideunes Edica Framework	2001	https://www.mlg	kotalwo-contentrisolo
excise integrate	organization	Malta Al Tankford	LODCL Hardework	2020	under Hansan Le B	THE WAY OF A CUT
Malta	Coventment	within the Office	Towards Trustworthy Al: Ethical A	2019	house (In site all	accommission/nab
Microsoft	Company	NA	Microsoft Responsible Al Principle		https://www.mis	
New Jealand	Covernment	Government Chi	Discussion Paper: International D		https://www.data	
New Zealand & Horiz	Multi-Org Group	N/8.	Reimagining Regulation for the A		tre front and	
06.00	Organization	K/A	Recommendation of the Council		https://legalitatr	
Secol.	Company	NA	OpenAl Charter		totas l'aper al co	
>	Multi-Org Group	N/A	Thematic Pillans			persiti pionial segulabeut
	Multi-Org Group		A Guide for Othical Data Science	2019	https://www.actu	
Singapore	Government		Discussion Paper on Artificial Inte		https://www.pdp	
Singapore	Coventmont		Model Al Governance Framework		https://www.pdp	
Singapore	Coventment		Principles to Promote Fairness, D		https://www.mas	Tracquareccy
CAE / Ocbal			Al Principles and Ethics for the Er		https://www.ama	
UNESCO & World Co	Multi-Org Droup		PRELIMINARY STUDY ON THE ETH		https://wresdoc.	
UNICEF	Organization		Policy guidance on Al for children Al and child rights policy		https://www.utic	
United Kingdom	Organization		Understanding artificial intelligen		https://www.turi-	
United Kingdom	Coventment		Interpretable Machine Learning		https://post.par5	
	Coventment		Cata Ethics Framework		https://www.gov	
			Ethical Principles for Artificial Inde	2020	https://www.ai.m	Indocrafitional_Particip
	Coventment	3440				
USA	Covernment Covernment		Artificial Intelligence Ethics Frame	2020	terrare (lowers) rite	Person continue priliki
USA USA			Artificial Intelligence Ethics Frame Pramoting the Use of Trustworth		https://www.intel	
USA USA USA	Covernment Covernment Covernment	Intel3 gence Corr	Promoting the Use of Trustworthy Othics Principles for Access to ano	2020	(https://trimpwt/	Transparent
USA USA USA USA USA	Covernment Covernment	Intelligence Com White House Veterar Affairs NSF & Amazon	Promoting the Use of Trustworth Dirics Principles for Access to ane NSE Program on Fairmess in Archiv	2020 2021 2021	(https://trimpwt/	Transparent A gewählter hitten de
USA USA USA USA USA	Covernment Covernment Covernment	Intelligence Com White House Veterar Affairs NSF & Amazon	Promoting the Use of Trustworthy Othics Principles for Access to ano	2020 2021 2021	https://trianged. https://www.olite https://www.cale	Transparent a gov/doct/sthica-ch Transparenzy
USA USA USA USA USA USA	Coventment Coventment Coventment Multi-Org Group Coventment Coventment	Intelligence Com White House Veteran Affains NSF & Amazon NSF - Council for NIST	Promoting the Use of Trustworthy Dition Principles for Access to ane NSF Program on Earness in Actifi 10 Simple Rules for Responsible I Draft Four Principles of Deplainab	2020 2021 2021 2027 2027 2020	https://trump.wt/ https://www.oit/ https://www.cit/ https://www.cit/ https://www.cit/	Distagarent a gevilteen ett ea da Distagaren ey las angle mener blai averatoren it en dee e
USA USA USA USA USA USA	Coveniment Coveniment Coveniment Multi-Org Group Coveniment Coveniment	Intelligence Com White House Veteran Affains NSF & Amazon NSF - Council for NIST NIST	Promoting the Use of Trustworthy Ethics Principles for Access to are NSF Program on Earness in Artifi- 10 Simple Rules for Responsible I Draft Four Principles of Explained U.S. LEADERSHIP IN AC A Plan for	2020 2021 2021 2021 2020 2020 2029	https://trumpwt. https://www.olf.c https://www.call. https://putties.c	Distagarent a gevilteen ett ea da Distagaren ey las angle mener blai averatoren it en dee e
USA USA USA USA USA USA USA	Coveniment Coveniment Coveniment Multi-Org Group Coveniment Coveniment Coveniment	Intelligence Com White House Veteran Atlain, NSF & Amazon NSF - Council for NIST NIST Office of Science	Promoting the Use of Trustworthy Ethics Principles for Access to ano NEP Program on Extrems in Ardia 10 Simple Rules for Presponsible I Draft Four Principles of Ecolands U.S. LEADERSHP IN Art A Francis AI Principles when drafting regula	2020 2021 2021 2027 2020 2020 2020 2020	https://www.oll. https://www.oll. https://www.oll. https://www.oll. https://www.oll. https://www.oll.	Distagarent a governer tetti ca da Pretaparency construction de open and transparent Disclosure and transp
USA USA USA USA USA USA USA USA USA USA	Coveniment Coveniment Multi-Org Droup Coveniment Coveniment Coveniment Coveniment Coveniment	Intelligence Com Rhite House Veteran Affain, NST 6 Amazon NST NST Office of Science Ad Hoc Committ	Promoting the Use of Trustworthy Erbics Principles for Access to an NSF Program on Fair tens in Arabi 10 Small Roberts for Responsible I Draft Four Principles of Explained U.S. LEARERSHEP IN AR A Francho Al Principles when drafting regula	2020 2021 2021 2020 2020 2020 2020 2020	International Action International Action International Action International Action International Action International Action International Action	a polytoport (rft) call do Transparency An analytoport (franchist) appended that More Declassive and transp Transparency
USA USA USA USA USA USA USA USA USA USA	Coveniment Coveniment Coveniment Multi-Org Snoup Coveniment Coveniment Coveniment Coveniment Coveniment Coveniment	Intelligence Com White House Veteran Affains NSF & Amazon NST NIST Office of Science Ad Hoc Committe White House CS	Promoting the Use of Trustworthy Ethics Principles for Access to an NSF Program on Eartesan in Artifi- to Simple Rules for Persponsible I Onth Four Principles of Explained U.S. LEADERSHIP IN ALLA Fain for All Principles when Stability regula Chait Agency Use of Artificial Intel American All Initiative Teaching	2020 2021 2021 2027 2020 2020 2020 2020	International Antonio State International An	Transporent services entre sa sea Transporency de and transporent open and transporent Discission and transport discission and transport discissi
USA USA USA USA USA USA USA USA USA USA	Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment	Intelligence Com Bilhite House Velenan Affain NSF & Amazon NSF Office of Science Ad Hoc Committe Bilhite House OS GSA	Premoting the Use of Trustworthy Dirics Principien for Access Is are NEP Program on Entress in Archi- I Carly Teach Principien of Explaines U.S. LEADERSEEP IN Arc A Plan for All Principies with charling regula- Carls Agency Use of Arthritical Intel American Al Initiative Year One Is Call Even President in All Efficient	2020 2021 2021 2027 2020 2029 2029 2020 2020 2020 2029 2029 2029	Integra (Internet of A Integra (Internet of A	Transparent augustatori etti basida Transparenty etti and di moranetti di su operi and transparent Discipare and transparent Discipare and transparent Casingarency operi transparenty operitri model develop
United Kingdom USA USA USA USA USA USA USA USA USA USA	Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment Coventment	Intelligence Com Bihlte House Veteran Aflam NSF & Amazon NSF - Council for NST Office of Science Ad Not Committ Bihlte House CSI GSA NJA	Promoting the Use of Trustworthy Ethics Principles for Access to an NSF Program on Eartesan in Artifi- to Simple Rules for Persponsible I Onth Four Principles of Explained U.S. LEADERSHIP IN ALLA Fain for All Principles when Stability regula Chait Agency Use of Artificial Intel American All Initiative Teaching	2020 2021 2021 2027 2020 2029 2029 2029 2020 2020 2029 2029 2029 2029 2029 2029	Inspan (Insert part) Inspan (Insert of La Inspan) (Insert of La	Transparent augustatori etti basida Transparenty etti and di moranetti di su operti and transparent Discipare and transparent Disciparenty casi etti dipercenti man operti model develop

ni ni nama na mana na

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"

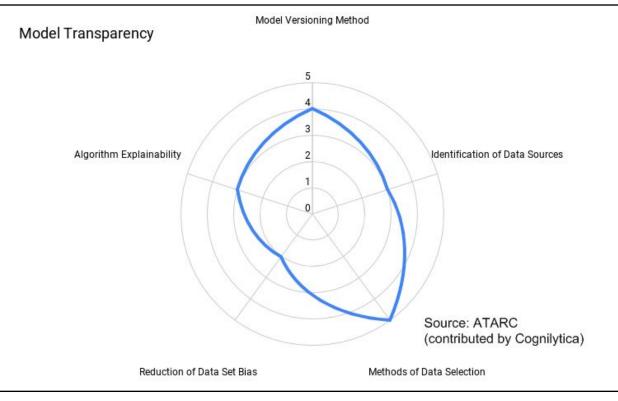
Completely absent of any systemic

transparency principles
```





# Assessing AI Transparency: Model Assessment Approach



www.cognilytica.co



# AI Governance: The Main Elements

## 🖈 🛛 System Auditability

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

## 🛨 Contestability

• Al systems need to provide ways to contest or appeal Al 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

## **r** 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

Sec Deschi Link

Rep Assessment & System Authobility Contemptoin

Jubercus/Dig Fremework/Whistie

Caurtry / Company Ora Tuter

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

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

credentialing or accountability

Dears Industry, St Al Ethics Formework 2019 billion i lower Rules on Responsible use of M JOIN TOTAL CHANNEL 1010 https://www. Ministry of Scient Development of responsible art Office of the Driv Drive of Account of the Dramason One Mind Ethics 2017 https://decorrind 2018 Stdatiwig JOSH STRAIN (Sector JOIP STREET, Delever.CO o Multi-Org Group N/H JOIF STRUGEN JOID BEES LINKS Microsoft Responsible & Prince 2019 https://www. Monetary Author Principles to Promote Fairness 6 Smart Dubai Gov Ar Principles and Ethics for the I Monking Group PRELIMITARY Shudy ON THE ETH 3019 TODA L'UNIO Ministry of Formy Policy guidance on Arthr childre 3030 https://www Office of Giotal 1-Al and child rights policy 3019 1121 ( max JOSH STREET Press JOJO NEED (Speed p. JOSO 1328 (hrat 2021 https://www 2020 https://www.fr Puli - Council for 12 Simple Rules for Responsible 2017 https://puntal Draft Faur Principles of Explained 2020 https://www.rl U.S. LENDERSHE'N ACA Plat for 2019 bits like Office of Science Al Principles when drafting regula 2020 https://www. 3020 1:322 ('www 2020 https://www.

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:
  - $\circ$  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"<sup>5</sup>
- 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"<sup>3</sup>
- "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?





#### 💟 @Cognilytica



## 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

One Type Subgroup / Ora Frammeark / Initiative Cor. Date" ( Link Organization EMERGING TECH Ethical Application of Artificial Int 2020 https://www. Ethical Framework for a Good ALS 2019 https://www.industry.gov.as/data 2018 http://mg.baidu.co Company Responsible Al The Monthial Declaration fo 2017 https://www. Canada 2013 https://www. Relieve Academy, Beiling Al Principles, ty Ethical Accountability Framew tigh-Level Exper Ethics Guidelines for Trustworthy Opinion of the Data Ethics Com-2019 https://www. Principles for responsible stewart 2020 https://epai.st/abox Google Al Principles 2018 https://si.google/principi Principled Artificial Intelligence mitarian Data S Organization A Framework for the Ethical Use ( Bennomible St. Key Thomas Co. The IEEE Global I ETHICALLY ALIENED DESIGN: A VI o Ethical Guidebres 2017 http://www.a Malta Al Taskics Microsoft Responsible Al Princip Company oric Multi-Org Group N/A NA OpenAl Charte 2018 https://openal.com/chatter Multi-Org Group N/A Invel Statistical Socia Multi-Ore Grean, N.W. Sinnipore Monetary Author Principles to Promote Fairness, E 2018 https://www.man.e HAT ( Dobie Smart Dubei Goy, Al Principles and Ethics for the Ex POST MANAGE STUDY ON THE FT. istry of Foreig Policy guidance on Al for childre Office of Global II Al and child rights policy Alam Turing Instit Understanding artificial intellias I K Darbarrard Di Internetable Marbine Learning Digi Data Dibica Framework Ethical Principles for Artificial Int Ethics Principles for Access to an Multi-Dre Group, NSE & Amazone, NSE Program on Extension Artific NST - Council for 10 Simple Dates for Support blat Draft Four Principles of Deplainab U.S. LEADERSHIP IN AL A Play In Office of Science Al Principles when drafting reguli Ad Hoc Commits Draft Ammey Use of Artificial Intel Coll Best Practices in Al Ethics Rome Call for Al Ethics



#### 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."

#### 🔰 @Cognilytica



## In Search of a Comprehensive Ethical AI Framework

## ... wait... we just presented it!

Societal Ethics	Responsible AI	Systemic AI Transparency	Al Governance	AI Explainability & Interpretability
<ul> <li>★ Human Values</li> <li>★ Al for Human Benefit</li> <li>★ Dignity</li> <li>★ Fairness</li> <li>★ Respect Diversity</li> <li>★ Avoid Bias</li> <li>★ Ensure Freedom &amp; Human Agency</li> <li>★ Keep the Human in Control</li> <li>★ Respect the Environment</li> </ul>	<ul> <li>★ Positive Purpose</li> <li>★ Law Abiding use of AI</li> <li>★ Ensure Safety</li> <li>★ Maintain Trust</li> <li>★ Provide Human Accountability</li> <li>★ Ensure Privacy</li> <li>★ No Lethal Autonomous Weapons</li> <li>★ Avoid Workforce Disruption</li> </ul>	<ul> <li>★ Visibility into AI Data</li> <li>★ Bias Measurement &amp; Mitigation</li> <li>★ Open Systems &amp; Human Decisions</li> <li>★ Provide Disclosure and Respect Consent</li> </ul>	Mitigation ★ System Auditability ★ Contestability	<ul> <li>★ Algorithmic Explainability</li> <li>★ Algorithmic Interpretability</li> <li>★ Root Cause Explanations</li> </ul>



# **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.

AI Systemic Transparency vs. Algorithmic Interpretability	
Another common concern with regards to Al systems is that humans are deleg authority to algorithms. Concerns about algorithmic decision-making are rele	
(R.	ir decisions through embled together to Igorithms to
Cognilytica Research 🛛 🛛 🥎	imic decisions.
State of Ethical AI	icci ated factors can help
Frameworks 2021	temic Al
Comparing, Contrasting, and Normalizing Concepts and	
Definitions of Evolving Frameworks for Ethical AI	
_	the c stes i Non-mateficence Privacy protection and securit
Doc. ID: CGR-ETH21	Protection of privacy and inter
April 27, 2021	
ABSTRACT	Privacy as a Human Right
The topic of chrics and responsibility comes up frequently in the context of Artificial instillighters, and for good reason. The concept of instighter machines is acry to many popic. Them any just as concented about the increasing power that humans have one data and the potential for malicious used powerful technology. To address these concents of a value of the concentration of the potential international and the potential for malicious used powerful technology. To address these concentrations are not acreated and the potential for malicious used the concentration of a value ways in which people can malinain control over the used potential. Nechnology in the valid excluding in a concent of excluding in concent and excluding in a concent data excluding in a concent of excluding in a concent of the potential structure of a concent of a cicital etics, responsibilite used PLA and years and potential weakness of different proposed ethical for the identified for tameworks and detertifies the strengths and potential weakness of different proposed ethical foremovics. Congritical as poposes a potential "momande" ethical foremovics that concepts identified in this document, with concise boundaries between thical concepts.	ens Bereit für spender Protogener fange Protogener fange
Al Contents Copyright 0 2021 Copylight. All rights reserved. Reproduction of this publication in any form without prior written permission is footdeten. The Information contained Herein has been obtained from sources believed: to be reliable. Copylight all disclaims all warmelines to be accuracy, comprises or advectary of unit homation. Copylight call how to in bibling for errors, entisistics or inadequasies in the information contained hereine or be interpretation thereo. The reader assumes sole responsibility for the accuracy, comprises where here in the interpretation thereo. The reader assumes sole responsibility for the accuracy comprises where here in the interpretation. The options expresses there are subject to charge without votion. All trademarks, endor take names are trademarked by their respective content and Copylifican takes on colliss to been annee.	or Internet & Society igital Innovation of al Council for ible use of Al in the marized below.
	bia Al Ethical Fra
Copyright © 2021 Copyhylca + Al Rights Reserved + www.copyhlylca.com + Info@copyilylca.com	stems must be ol over decision
Bias / Discrimination     Ono-discrimination: Al cannot have result	s or responses that threaten
the well being of a specific group or limit th	he rights of historically
marginalized populations. There must be c	constant monitoring of false





# **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

Or Organizations are increasingly making use of A systems to power their operations and emable a wide range of applications from the invitial to the missions chick. If ways of these systems are being put time applications that the initial power initians of the systems in a system are being put the systems are being put to applications that the put the systems is made put the systems in which are provide grader that, provide visibility and transport, and utilizes not provide systems are being put the systems. Further, provide grader the systems is made that can provide grader that, and provide systems are deviced by a systems, and utilizes the systems. Unfortunately, many of the existing relation and non-profit granupare and the relation and frameworks has that can serve as a basis for individual organizations all non-profits graners and inspitzed for relations for an implementation or sense in singization for relations and inspitzed for relations and memory sense. The systems is under the systems in the systems in the system are being as a template for relation at a many sense of the systems in the systems. Sense of the systems is the system are as a basis for individual organizations at a many sense of the systems. Sense of the systems is the system are as a basis for individual organizations at the systems. Sense of the systems is the systems are as a basis for individual organizations at the systems. Sense of the systems are as a basis for individual organizations at the systems. Sense of the systems are as a basis for individual organizations at the systems. Sense of the systems are as a basis for the sys

All Contents Capyright 2: 2012 Cognitypics. This decument is relaxed under Crashive Common Mathemises. A Common Common Common Common Common Common Relaxed and Common Common Common Common Common Common Common Relaxed and Common Common Common Common Common Common Common Relaxed Common Commo

vyright © 2021 Cognilytica • CC BY-NC-SA 4.0 • www.cognilytica.com • info@cognilytica.com





# Thank You

**Presenters:** 

Kathleen Walch, Ronald Schmelzer, Megan Yarbrough

Cognilytica - http://www.cognilytica.com

info@cognilytica.com



# **Appendix / Sources**



## Sources

- 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. <u>https://doi.org/10.3390/electronics8080832</u>
- 4. <u>https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48</u>
- 5. "Error-riddled data sets are warping our sense of how good AI really is" -<u>https://www.technologyreview.com/2021/04/01/1021619/ai-data-errors-warp-machine-learning-prog</u> <u>ress/</u>

