Technical Advisory Council Meeting

October 8, 2020

THELINUX FOUNDATION



Antitrust Policy Notice

- Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.
- Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.





Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the TAC Wiki





Reminder: LF AI Useful Links

Web site: <u>lfai.foundation</u>

Wiki: <u>wiki.lfai.foundation</u>

GitHub: github.com/lfai

Landscape: <u>landscape.lfai.foundation</u> or <u>l.lfai.foundation</u>

Mail Lists: https://lists.lfai.foundation

LF Al Logos: https://github.com/lfai/artwork/tree/master/lfai

LF AI Presentation Template:

https://drive.google.com/file/d/leiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing

Events Page on LF AI Website: https://lfai.foundation/events/

Events Calendar on LF Al Wiki (subscribe available):

https://wiki.lfai.foundation/pages/viewpage.action?pageId=12091544

Event Wiki Pages: https://wiki.lfai.foundation/display/DL/LF+AI+Foundation+Events





Agenda

- > Roll Call (3 mins)
- Approval of Minutes (3 mins)
- Welcome Herron Tech as General Member!
- Trusted Al Project Presentations
 - Adversarial Robustness Toolbox (12 minutes)
 - Al Fairness 360 (12 minutes)
 - Al Explainability 360 (12 minutes)
- Upcoming TAC Meetings (5 minutes)
- > LF AI General Updates (5 minutes)
- Open Discussion (10 minutes)

>





TAC Voting Members

Board Member	Contact Person	Email		
AT&T	Anwar Atfab	anwar@research.att.com		
Baidu	Daxiang Dong	dongdaxiang@baidu.com		
Ericsson	Rani Yadav-Ranjan	rani.yadav-ranjan@ericsson.com		
Huawei	Huang Zhipeng	huangzhipeng@huawei.com		
IBM	Susan Malaika	malaika@us.ibm.com		
Nokia	Jonne Soininen	jonne.soininen@nokia.com		
Tech Mahindra	Nikunj Nirmal	nn006444@techmahindra.com		
Tencent	Bruce Tao	brucetao@tencent.com		
Zilliz	Jun Gu	jun.gu@zilliz.com		
ZTE	Wei Meng	meng.wei2@zte.com.cn		
Graduate Project	Contact Person	Email		
Acumos	Nat Subramanian	natarajan.subramanian@techmahindra.com		
Angel	Bruce Tao	brucetao@tencent.com		
Horovod	Travis Addair	taddair@uber.com		
ONNX	Jim Spohrer (Chair of TAC)	spohrer@us.ibm.com		

Approval of Sept 24th TAC Minutes

Draft minutes from the September 24th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

That the minutes of the September 24th meeting of the Technical Advisory Council of the LF Al Foundation are hereby approved



Welcome!

Herron Tech became a General Member of Linux Foundation Artificial Intelligence Foundation on September 25, 2020.

Herron Tech





Invited Project Presentations Adversarial Robustness Toolbox (Matthieu Sinn) Al Fairness 360 (Kush Varshney) Ai Explainability 360 (Michael Hind)





Adversarial Robustness Toolbox (ART) – Evasion, Poisoning, Extraction and Inference



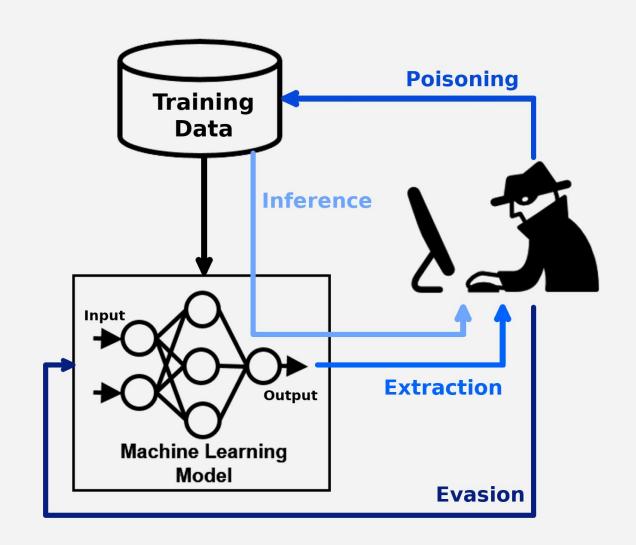
October 2020



Adversarial Threats

Adversarial threats against machine learning models and applications have a wide variety of attack vectors.

- **Evasion:** Modifying input to influence model
- Poisoning: Modify training data to add backdoor
- Extraction: Steal a proprietary model
- Inference: Learn information on private data



Evasion.

Imperceptible modifications to medical images to influence classification.

Poisoning.

Imperceptible patterns in training data create backdoors that control models.

Extraction.

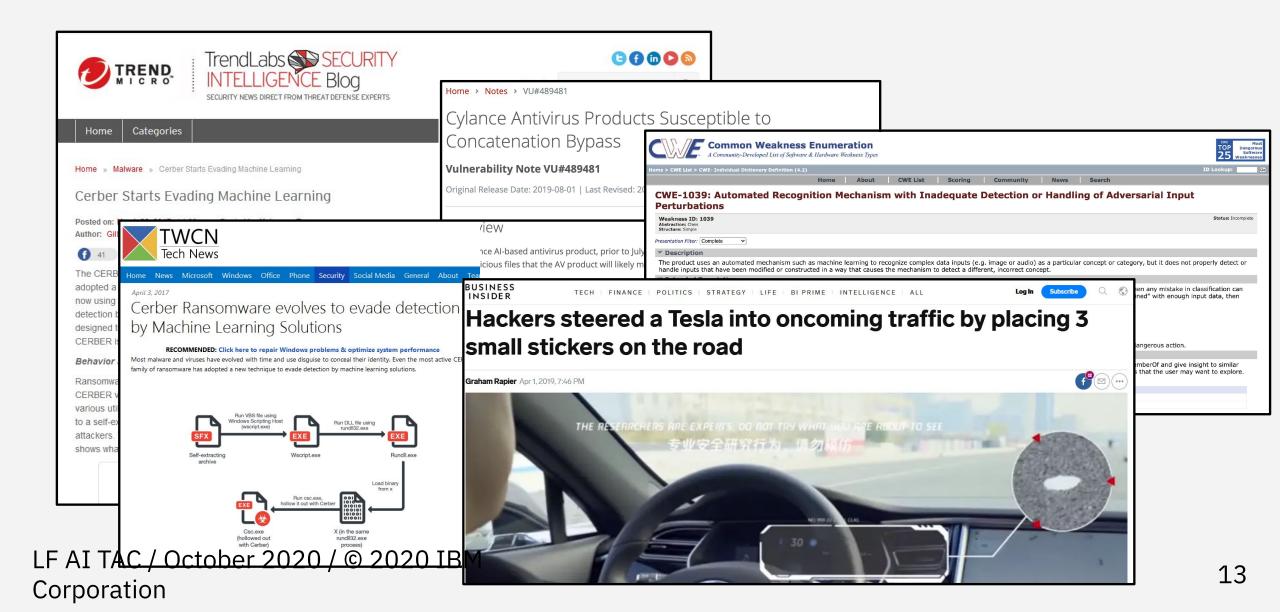
Theft of proprietary models through model queries.

Inference.

Derive properties of the model's training data up to identifying single data entries.

LF AI TAC / October 2020 / © 2020 IBM Corporation

Real Adversarial Threats



Adversarial Threat Combinations

Combinations of adversarial threats become more effective than their sum.

- Extraction attacks enable stronger white-box evasion attacks
- Extraction attacks steal models that could leak more private information in inference attacks



LF AI TAC / October 2020 / © 2020 IBM Corporation

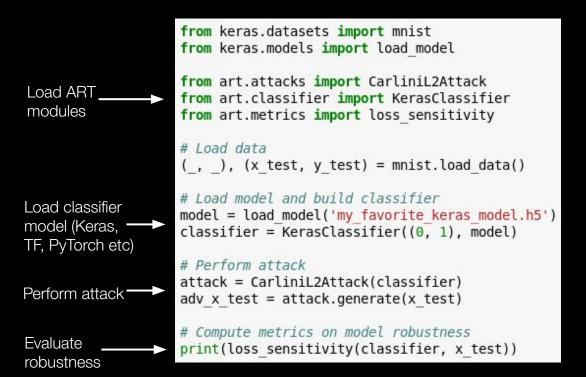
Adversarial Robustness Toolbox (ART)

Repo: https://github.com/Trusted-Al/adversarial-robustness-toolbox

Docs: https://adversarial-robustness-toolbox.readthedocs.io/

Demo: https://art-demo.mybluemix.net

- Python library, 14K lines of code
- State-of-the-art <u>attacks</u>, <u>defences</u> and <u>robustness metrics</u>





K Keras



PYTORCH





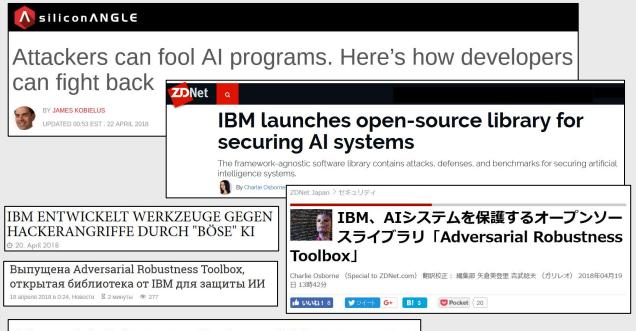
LightGBM

Open-source release @ RSA 2018:



Current stats:

- 1.8K GitHub stars
- 500+ forks
- 250+ clones/w
- 1K+ downloads/w



Adversarial Robustness Toolbox : IBM propose une boite à outils open source pour sécuriser l'intelligence artificielle

23-04-2018 | door: Witold Kepinski



IBM Adversarial Robustness Toolbox beschermt tegen

kwaadaardige Al



ART is a Python library for machine learning security.







GPy



LightGBM



CatBoost

LF AI TAC / October 2020 / © 2020 IBM Corporation

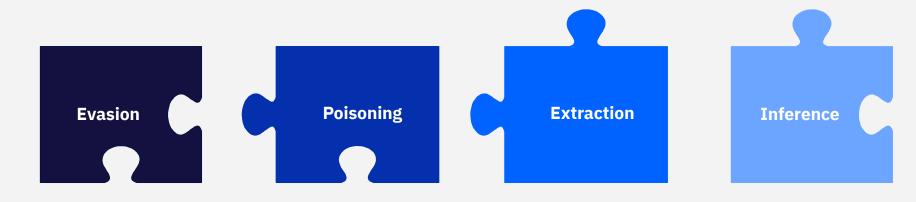
- github.com/Trusted-AI/adversarial-robustness-toolbox
- 500+ forks, 1.8K stars
- providing tools to developers and researcher
- Evaluating and Defending machine learning models and applications
- All Tasks: Classification, Object Detection, Automated Speech Recognition, etc.
- All Frameworks: TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy
- All Data: images, tables, audio, video, multi-modal, etc.

16

The Tools of ART

ART 1.3/1.4

- art.metrics
 - Methods to quantify robustness
- art.estimators
 - Abstractions for models



art.attacks examples	 21 (+8) White-box (e.g. FGSM, PGD, Carlini&Wagner,) Black-box (HopSkipJump, Boundary, ZOO,) 	 4 (+2) Backdoor, Feature Collision, SVM, Adversarial Embedding 	 3 FunctionallyEquivalent KnockOffNets CopyCat	 6 (+6) Model Inversion Attribute Inference Membership Inference
art.defences examples	 15 (+4) Adversarial Training (Madry, Fast is Better than Free,) Preprocessing Transformer 	 5 (+2) Detection (Activation, Provenance, RONI, Spectral Signature,) Transformation (Neural Cleanse) 	 6 Postprocessing (Reverse Sigmoid,) 	Differential Privacy Library
LF AI TAC / October 2020 / Corporation	©·2020 IBM			17

The Roadmap for ART

v0.1 (Apr 2018): v0.3 (Aug 2018): v1.0 (Sep 2019): v1.1 (Jan 2020): v1.3 (Jun 2020): v1.4 (Sep 2020): - non-DL models extraction attacks - non-classification - automated speech - image classification - data poisoning - evasion attacks - non-image inputs tasks attacks recognition - DL models - object detection - inference attacks

v1.5 (Dec 2020):

- extend ASR capabilities
- extend support for multi-modal inputs
- add SOTA attacks & baseline defenses

v1.6 (Mar 2021):

- support for physical-world threat models
- support models with discrete inputs (e.g. text)
- develop higher-level / composite evaluations

... and beyond:

- use cases (cybersecurity, forensics etc.)
- end-to-end test suites
- novel tasks (regression, RL learning etc.)
- keep up with SOTA attacks & defenses
- improve performance and usability

The Roadmap for ART

v0.1 (Apr 2018): v0.3 (Aug 2018): v1.0 (Sep 2019): v1.1 (Jan 2020): v1.3 (Jun 2020): v1.4 (Sep 2020): - non-DL models - extraction attacks - non-classification - automated speech - image classification - data poisoning - evasion attacks - non-image inputs tasks recognition attacks - DL models - object detection - inference attacks

v1.5 (Dec 2020):

- extend ASR capabilities
- extend support for multi-modal inputs
- add SOTA attacks & baseline defenses

v1.6 (Mar 2021):

- support for physical-world threat models
- support models with discrete inputs (e.g. text)
- develop higher-level / composite evaluations

... and beyond:

- use cases (cybersecurity, forensics etc.)
- end-to-end test suites
- novel tasks (regression, RL learning etc.) - keep up with SOTA attacks & defenses
- improve performance and usability

External contributions / collaborations particularly welcome!

Thank you

Beat Buesser, Mathieu Sinn IBM Research Europe Dublin, Ireland

The Adversarial Robustness Toolbox (ART) Authors 2020 http://github.com/Trusted-AI/adversarial-robustness-toolbox

AI Fairness 360

Kush R. Varshney Distinguished Research Staff Member and Manager



AI is powering critical workflows and trust is essential









loan processing

employment

customer management

quality control

Multiple factors are placing fairness of AI as a top priority



brand reputation

increased regulation

focus on social justice

Unwanted bias

places privileged groups at systematic advantage IBM Research / October 8, 2020 / © 2020 IBM Corporation

and unprivileged groups at systematic disadvantage. IBM Research / October 8, 2020 / © 2020 IBM Corporation

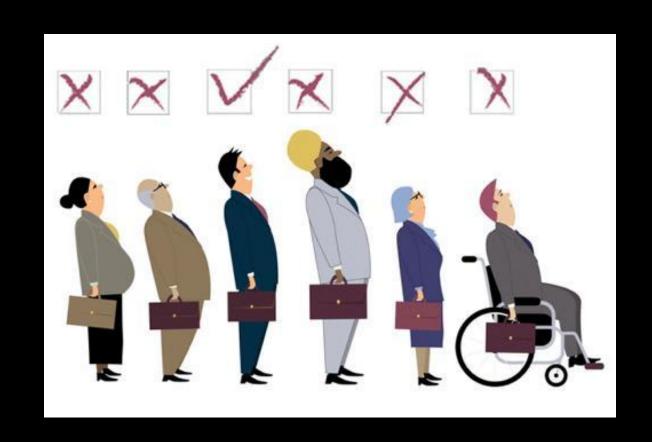
Where does unwanted bias come from?

Problem misspecification.

Data engineering.

Prejudice in historical data.

Undersampling.



The most comprehensive toolkit for handling bias in machine learning.

http://github.com/trusted-ai/aif360

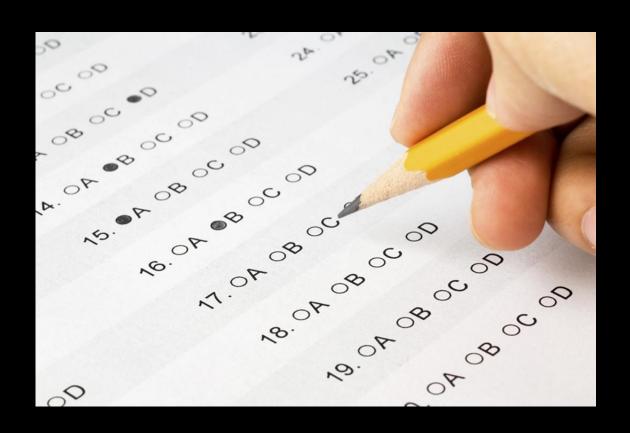
- R and Python
 - Python version transitioning to full scikit-learn compatibility
- Comprehensive set of fairness metrics
 - Group fairness
 - Individual fairness
- 11 state-of-the-art bias mitigation algorithms
- http://aif360.mybluemix.net
 - Extensive industry tutorials to educate users and practitioners
 - Interactive demo

IBM Research / October 8, 2020 / © 2020 IBM Corporation

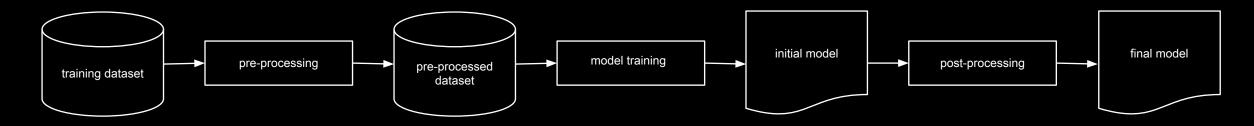
There are many ways to measure bias

Example:

statistical parity vs. equal opportunity



There are many ways to mitigate bias

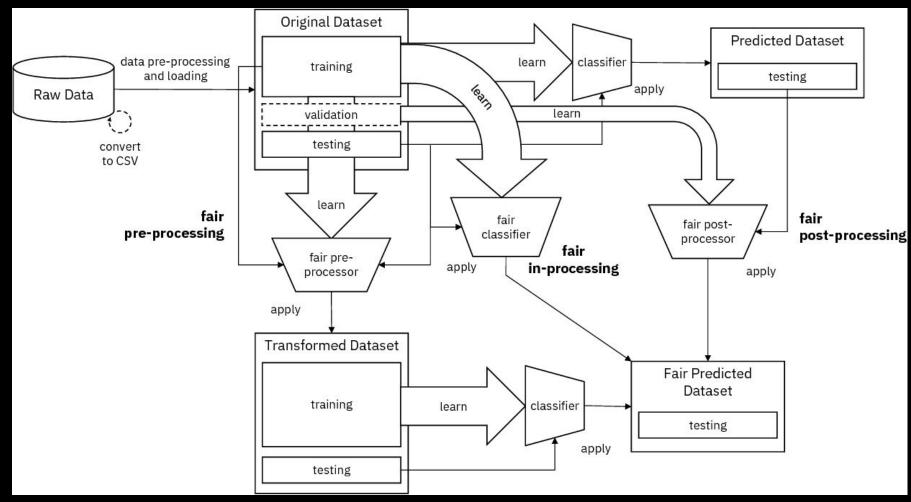


bias mitigation pre-processing

bias mitigation in-processing

bias mitigation post-processing

Bias mitigation algorithms



IBM Research / October 8, 2020 / © 2020 IBM Corporation

- -Full support for scikit-learn compatibility
- -Full coverage of R version
- -Pre-processing, in-processing, and post-processing bias mitigation algorithms for graphs
- -Greater support for individual fairness and text
- Domain-specific extensions
 - Social justice



Thank you

Kush R. Varshney Distinguished Research Staff Member and Manager

krvarshn@us.ibm.com

Explainability 360

LF AI TAC October 8, 2020

Michael Hind Distinguished Research Staff Member IBM Research, Yorktown Heights

@michael_hind

The Call for Explainability

CIO JOURNAL.

Companies Grapple With AI's Opaque Decision-Making Process
THE WALL STREET JOURNAL

Can A.I. Be Taught to Explain Itself?

The New York Times Magazine

When a Computer Program Keeps You in Jail

Criteria for parole algorithm was not available to parolee.

Why Explainable AI Will Be the Next Big Disruptive Trend in Business AlleyWatch

This field of XAI is going to be hugely important, with a number of important social, legal and ethical implications.

"Capital One ... would like to use deep learning for all sorts of functions, including deciding who is granted a credit card. But **it cannot do that because the law requires companies to explain the reason** for any such decision to a prospective customer."

MIT TR, Apr, 2017

"The agency (CIA) cannot just be accurate, it's also got to be able to demonstrate how it got to the end result. So if an analytic isn't explainable, it's not "decision-ready."

Defense One, June 2019

Society is starting to require explanations but what does it mean?

The General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing and profiling (Art.22)

Paul Nemitz, Principal Advisor, European Commission Talk at IBM Research, Yorktown Heights, May, 4, 2018

?

Illinois and City of Chicago Poised to Implement New Laws Addressing Changes in the Workplace – Signs of Things to Come? (US)

Wednesday, June 5, 2019

Illinois Restricts Use of Artificial Intelligence in Hiring

On May 29, 2019, the Illinois Legislature unanimously passed the Artificial Intelligence Video Interview Act, which, not surprisingly, addresses how employers use artificial intelligence to analyze job applicant video interviews to determine the applicant's fitness for the position. Under the new law (assuming it is signed by the Governor, as anticipated), before requesting an applicant submit to a video interview, employers will be required to:

- notify applicants for positions based in Illinois that it plans to have their video interview analyzed electronically;
- explain how the artificial intelligence analysis technology works what general characteristics it will use to evaluate candidates; and
- obtain the applicant's consent to these procedures (note: consent does not have to be in writing).

NATIONAL LAW REVIEW

There Exists a Large Research Community Advancing AI Explainability

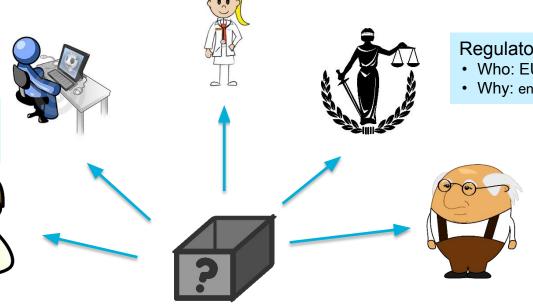


The Fifth Annual Workshop on Human Interpretability in Machine Learning (WHI 2020), held in conjunction with ICML 2020, will bring together artificial intelligence (AI) researchers who study the interpretability of AI systems, develop interpretable machine learning algorithms, and develop methodologies to interpret black-box machine learning models (e.g., post-hoc interpretations). This is a very exciting time to study interpretable machine learning, as the advances in large-scale optimization and Bayesian inference that have enabled the rise of black-box machine learning are now also starting to be exploited to develop principled approaches to large-scale interpretable machine learning. Interpretability also forms a key bridge between machine learning and other AI research directions such as machine reasoning and planning.

Meaningful Explanations Depend on the Explanation Consumer

End Users

- Who: Physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights(?)



Regulatory Bodies

- Who: EU (GDPR), NYC Council, US Gov't, etc
- Why: ensure fairness for constituents

Affected Users

- Who: Patients, accused, loan applicants, teachers
- Why: understanding of factors

Must match the complexity capability of the consumer Must match the domain knowledge of the consumer

© 2018 IBM Research AI

Al System builders, stakeholders

Why: ensure/improve performance

• Who: data scientists, developers, prod mgrs

IBM AI Explainability 360

The most comprehensive **open source** toolkit for explaining ML models and data:

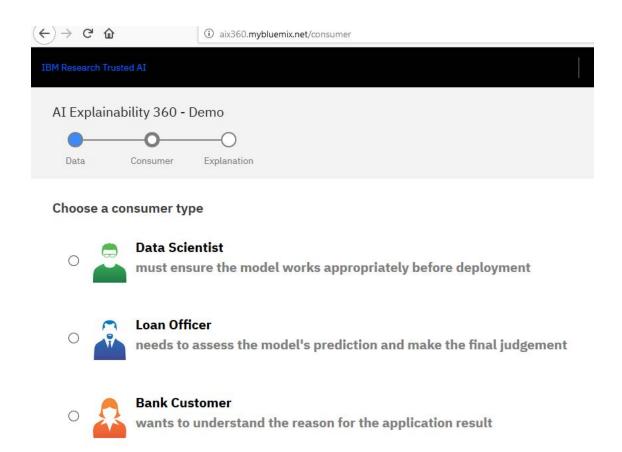
- 8 innovated algorithms from IBM Research + 2 other popular algorithms
- An interactive demo that provides a gentle introduction through a credit scoring application
- 13 tutorial notebooks covering use cases in finance, healthcare, lifestyle, retention, etc.
- documentation that guides the practitioner on choosing an appropriate explanation method.

One Explanation Does Not Fit All:

A Toolkit and Taxonomy of Al Explainability Techniques

by Arya et al.

https://arxiv.org/abs/1909.03012

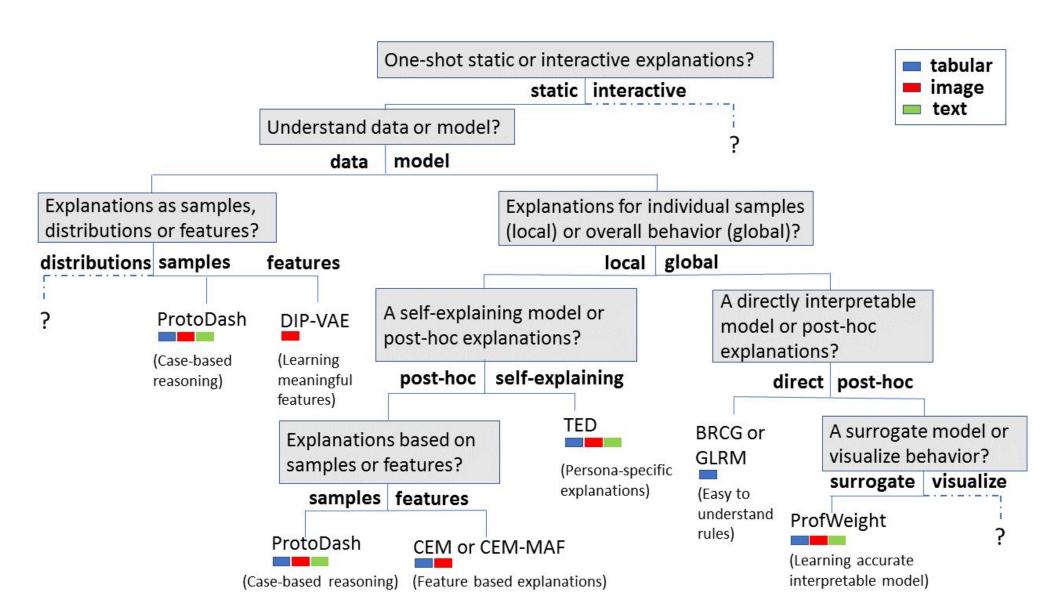


http://aix36o.mybluemix.net/

One Explanation Does Not Fit All: A Toolkit and Taxonomy of Al Explainability Techniques

by Arya et al.

https://arxiv.org/abs/1909.03012



Impact

Caveat: We only know what we hear ...

- Company 1: used toolkit to formulate and create an enterprise-wide center of competency in explainability
- US regulator: using toolkit to learn space and guide regulation
- FICO Explainability Challenge winner
- Improve model accuracy in Semiconductor Manufacturing
- Improve adoption/trust in IT customer care model

Metric	Value
Forks	144
Stars	648
Avg unique clones/day	2.9
Avg visits/day	145
Avg PyPi downloads/month	823
Slack users	190
Closed PRs	62
Tutorial views	4,023

- "What a fantastic resource (AIX360 is)! Thanks to everyone working on it." John C. Havens, Executive Director of IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- "I have found aix360 to be most comprehensive." Arpit Sisodia, Data Scientist with Ericsson

Desirable Contributions

AIX360 is the most comprehensive open source explainability toolkit, but it can be even more comprehensive

- Contributions for missing categories in Taxonomy
 - "interactive"
 - Static, data, distributions
 - Static, model, global, post-hoc, visualize: significant publications exist
- Support for missing modalities (e.g., contrastive for text)
- New categories not in taxonomy
- Support for the diversity of deep-learning frameworks
- Framework-specific model classes that expose a common API needed by explainability algorithm developers.

Further Details

One Explanation Does Not Fit All:
A Toolkit and Taxonomy of Al Explainability Techniques

by Arya et al. https://arxiv.org/abs/1909.03012

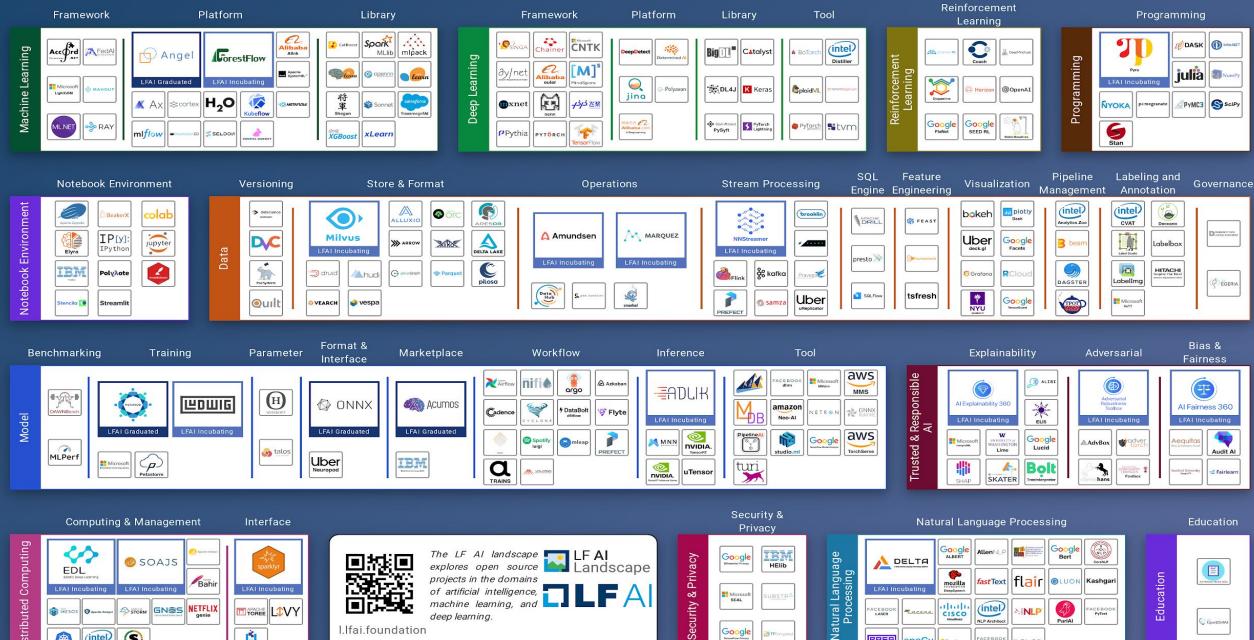
LF AI General Updates





See the interactive landscape at I.Ifai.foundation

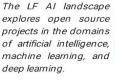
Greyed logos are not open source



















Q OpenDS4A

Graduated LFAI Projects (4)









A Growing LF Al Project Portfolio and Community



Incubating LFAI Projects (15)



























LF Artificial Intelligence Foundation



Companies hosting projects in LF AI

https://landscape.lfai.foundation/format=hosting









































Looking to host a project with LF AI

Hosted project stages and life cycle:

https://lfai.foundation/project-stages-and-lifecycle/

Offered services for hosted projects:

https://lfai.foundation/services-for-projects/

Contact:

Jim Spohrer (TAC Chair) and Ibrahim Haddad (ED, LF AI)





Promoting Upcoming Project Releases

We promote project releases via a blog post and on LF AI <u>Twitter</u> and/or <u>LinkedIn</u> social channels

For links to details on upcoming releases for LF AI hosted projects visit the <u>Technical</u> <u>Project Releases wiki</u>

If you are an LF AI hosted project and would like LF AI to promote your release, reach out to pr@lfai.foundation to coordinate in advance (min 2 wks) of your expected release date.





Note on quorum

As LF AI is growing, we now have 14 voting members on the TAC.

TAC representative - please ensure you attend the bi-weekly calls or email Jacqueline/Ibrahim to designate an alternate representative when you can not make it.

We need to ensure quorum on the calls especially when we have items to vote on.



Updates from the Outreach Committee

Events

- > Upcoming Events
 - > Visit the LF Al Events Calendar or the LF Al 2020 Events wiki for a list of all events
 - To participate visit the <u>LF AI 2020 Events wiki page</u> or email info@lfai.foundation

- > Please consider holding virtual events
 - To discuss participation, please email events@lfai.foundation



Upcoming Events

LF Al Booth at OSS EU – Oct 26-28

October 26 - October 28

LF AI Foundation will have a booth at Open Source Summit Europe (OSS EU)

"AI/ML/DL presented by LF AI Foundation" Track at OSS EU – Oct 26-28

October 26 - October 28

"AI/ML/DL presented by LF AI Foundation" Track at Open Source Summit Europe (OSS EU)

LF Al Mini Summit at OSS EU - Virtual - Oct 29

October 29 @ 12:00 am

LF AI Foundation will hold a Mini Summit at Open Source Summit EU (OSS EU)





LF AI PR/Comms

- Please follow LF AI on <u>Twitter</u> & <u>LinkedIn</u> and help amplify news via your social networks Please retweet and share!
 - > Also watch for news updates via the tac-general mail list
 - View recent announcement on the <u>LF Al Blog</u>

 Open call to publish project/committee updates or other relevant content on the LF Al Blog

 To discuss more details on participation or upcoming announcements, please email pr@lfai.foundation



Call to Participate in Ongoing Efforts

Trusted Al

Leadership:

Animesh Singh (IBM), Souad Ouali (Orange), and Jeff Cao (Tencent)

- Goal: Create policies, guidelines, tooling and use cases by industry
- **Github:**https://github.com/lfai/trusted-ai
- Wiki: https://wiki.lfai.foundation/display/DL/Trusted+AI+C ommittee
- > To participate: https://lists.lfai.foundation/g/trustedai-committee/
- Next call: Bi-weekly on Thursdays at 7am PT, subscribe to group calendar on wiki https://wiki.lfai.foundation/pages/viewpage.action?pageld=12091895

ML Workflow & Interop

- Leadership:
 - Huang "Howard" Zhipeng (Huawei)
- Goal:
 Define an ML Workflow and promote cross project integration
- Wiki: https://wiki.lfai.foundation/display/DL/ML+Workflo w+Committee
- > To participate: https://lists.lfai.foundation/g/mlworkflow-committee
- Next call: Every 4 weeks on Thursdays at 7:00 am PT, subscribe to group calendar on wiki https://wiki.lfai.foundation/pages/viewpage.action?pageld=18481242





Launching an effort to create Al Ethics Training

Initial developed course by the LF: Ethics in Al and Big Data - published on edX platform:

https://www.edx.org/course/ethics-in-ai-and-big-data

The goal is to build 2 more modules and package all 3 as a professional certificate - a requirement for edX

To participate:
 https://lists.lfai.foundation/g/aiethi
 cs-training





Upcoming TAC Meetings





Upcoming TAC Meetings

- October 22: Saleforce Data Modeling Discussion
- November 5: TBD

Please send agenda topic requests to tac-general@lists.lfai.foundation



TAC Meeting Details

- > To subscribe to the TAC Group Calendar, visit the wiki: https://wiki.lfai.foundation/x/XQB2
- Join from PC, Mac, Linux, iOS or Android: https://zoom.us/j/430697670
- Or iPhone one-tap:
 - US: +16465588656,,430697670# or +16699006833,,430697670#
- Or Telephone:
 - > Dial(for higher quality, dial a number based on your current location):
 - US: +1 646 558 8656 or +1 669 900 6833 or +1 855 880 1246 (Toll Free) or +1 877 369 0926 (Toll Free)
- Meeting ID: 430 697 670
- International numbers available: https://zoom.us/u/achYtcw7uN



Open Discussion





Legal Notices

- The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at https://www.linuxfoundation.org/trademark-usage, as may be modified from time to time.
- Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at https://lmi.linuxfoundation.org for details regarding use of this trademark.
- > Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.
- > TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.
- Facebook and the "f" logo are trademarks of Facebook or its affiliates.
- LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.
- YouTube and the YouTube icon are trademarks of YouTube or its affiliates.
- All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.
- The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at https://www.linuxfoundation.org/privacy and its Antitrust Policy at https://www.linuxfoundation.org/antitrust-policy. each as may be modified from time to time. More information about The Linux Foundation's policies is available at https://www.linuxfoundation.org.
- > Please email legal@linuxfoundation.org with any questions about The Linux Foundation's policies or the notices set forth on this slide.

