On the need of an Ethical AI Due Diligence

Prof. Roberto V. Zicari
Frankfurt Big Data Lab
www.bigdata.uni-frankfurt.de

June 28, 2019
Data Research Centre, Campus Fryslân
and DataFryslân,
The Netherlands
AI is Here.
Deep learning architectures such as deep neural networks, recurrent neural networks, and convolutional neural networks have been applied to fields including:

- Computer Vision
- Speech Recognition
- Natural Language Processing
- Audio recognition,
- Medical image analysis, …

where they have produced results comparable to and in some cases superior to human experts. Source: Wikipedia

Siri, Alexa, Tesla, Amazon.com, Netflix, Smart Phones, Social Media…
AI (in some cases) produces superior results to human experts.

“AI beats expert doctors at finding cervical pre-cancers”

60,000 cervical images collected from Costa Rica.
The study began in the 1990s, involving more than 9,400 women who were followed for up to 18 years.
AI: 91 percent accuracy vs. Human: 71 percent
National Cancer Institute's Division of Cancer Epidemiology and Genetics Washington, USA
“...our application of mobile and sensor technology to monitor symptoms, disease progression and treatment response – the so called “Digital Biomarkers”.

We have our most advanced programmes in Multiple Sclerosis (MS) and Parkinson`s Disease (PD), with several more in development. Using these tools, a longitudinal real-world profile is built that, in these complex syndromes, helps us to identify signals and changes in symptoms or general living factors, which may have several potential benefits.”

– Bryn Roberts  
Global Head of Operations for Roche Pharmaceutical Research & Early Development

The individual and collective conscience is the existential place where the most significant things happen.

- Awareness allows us to live a life, not react to it…
- Awareness happens, but needs to be continuously nurtured.
- If I am aware, am I also conscious?
- Awareness is considered a prerequisite for consciousness. You can’t be conscious about something if you’re unaware of it.
- Consciousness is a choice we can make (or not).

Source: http://www.bigdata.uni-frankfurt.de/ethics-artificial-intelligence/
What Do We Worry About?

Do we worry about AI?

Do we worry about the People developing AI?

Do we worry about the People deciding to use AI?

Do we worry about the People using AI?

We do not worry
“You cannot inspect what is absent”

– Gio Wiederhold

Do no harm
Can we explain decisions?

What if the decision made using AI-driven algorithm harmed somebody, and you cannot explain how the decision was made?

At present we do not really understand how Advanced AI-techniques such as used in Deep learning (e.g. neural networks) really works. It can be extremely difficult to understand which features (millions of features) of the data the machine used, and how they were weighted, to contribute to the outcome.

This is due to the technical complexity of such advanced neural networks, which need huge amount of data to learn properly. It is a try and error.

This poses an ethical and societal problem.
No absolute True Results: Probability

Source: https://slideplayer.com/slide/5098348/
Practically: When do we harm?

Accuracy

“What happens if my algorithm is wrong? Someone sees the wrong ad. What’s the harm? It’s not a false positive for breast cancer.” (*)

-- Claudia Perlich, Data Scientist

But Marketing/Social Media are not really harmless. …and we do have fake news! We also have false negative…

(*) Source: Big Data and The Great A.I. Awakening. Interview with Steve Lohr ODBMS Industry Watch, December 19, 2016
False Positive, False Negative

- A **false positive** is an error in data reporting in which a test result improperly indicates presence of a condition, such as a disease (the result is positive), when in reality it is not present.

- A **false negative** is an error in which a test result improperly indicates no presence of a condition (the result is negative), when in reality it is present.

*Source: Wikipedia*
Bias

When algorithms are used for example, to review loan applications, recruit new employees or assess potential customers, if the data are skewed the decisions recommended by such algorithms may be discriminatory against certain categories or groups.

Technically (*) Bias in machine learning = errors in estimation or over/under representing populations when sampling.

Selection, sampling, reporting bias
Bias of an estimator
Inductive bias

Other kinds of bias (**)  
Allocative harm = when a system allocates or withholds a certain opportunity or resource

Representation harm = when a system reinforces the subordination of some groups along the lines of identity

(*) Source: CS 294: Fairness in Machine Learning UC Berkeley, Fall 2017 https://mrtz.org/nips17/#/6

(**) Source: Kate Crawford, Keynote “The Trouble with Bias” Neural Information Processing System Conference
Suppose two people are tasked with developing a system to sort a basket of fruit. They have to determine which pieces are “high quality” and will be sold at the market, and which will instead be used for making jam.

Both people are given the exact same data — the fruit — and the same task... Given the same task and data, the two people are likely to have different results.

Perhaps one person believes the primary indicator of a fruit’s quality is brightness of color. That person may sort the fruit based on how vibrant it is, even though not all fruits are brightly colored; that person would send strawberries to the market and melons to the jam factory. Meanwhile, the other person might believe that unblemished fruit is the best quality, even though fruits with protective rinds might look scruffy on the outside, but are perfectly fine on the inside; that person could send unripe strawberries to the market and ripe melons or bananas to the jam factory.

These different, yet similarly logical and evenly applied criteria, will result in two different outcomes for the same basket of fruit.”

ALGORITHMIC CONSEQUENCES
It’s one thing to have an algorithm that marginalizes melons or unfairly sorts cucumbers, but what happens when algorithms make important decisions about humans?

Source: Understanding Bias in Algorithmic Design
https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e
Diversity

“If AI/ML teams are too homogeneous, the likelihood of group-think and one-dimensional perspectives rises – thereby increasing the risk of leaving the whole AI/ML project vulnerable to inherent biases and unwanted discrimination." -- Nicolai Pogadl (*)

(*) Source: personal communication.
“Deep neural networks can fail to generalize to out-of-distribution inputs, including natural, non-adversarial ones, which are common in real-time settings”. (*)

(*) Source: **Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects** Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, Anh Nguyen (Submitted on 28 Nov 2018 (v1), last revised 13 Jan 2019 version, v2)
Several machine learning models, including neural networks, consistently misclassify adversarial examples---inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. (**) 

(**) Source: Explaining and Harnessing Adversarial Examples Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy (Submitted on 20 Dec 2014 (v1), last revised 20 Mar 2015 version, v3)
Let's consider an autonomous car that relies entirely on an algorithm that had taught itself to drive by watching a human do it.

What if one day the car crashed into a tree, or even worse killed a pedestrian?
The Uber Case for *False positive for plastic bags*…

„The newsletter "The Information" has reported a leak from Uber about their fatal accident. The relevant quote:

The car’s sensors detected the pedestrian, who was crossing the street with a bicycle, but Uber’s software decided it didn’t need to react right away. **That’s a result of how the software was tuned.** Like other autonomous vehicle systems, Uber’s software has the ability to ignore “false positives,” or objects in its path that wouldn’t actually be a problem for the vehicle, such as a plastic bag floating over a road. In this case, Uber executives believe the company’s **system was tuned so that it reacted less to such objects.** But the tuning went too far, and the car didn’t react fast enough, one of these people said.“ (*)

(*) How reliable is this Source? : [https://ideas.4brad.com/uber-reported-have-made-error-tuning-perception-system](https://ideas.4brad.com/uber-reported-have-made-error-tuning-perception-system)

Story also in Der Spiegel Nr. 50/8.12.2018 *Tod durch Algorithms* (Philipp Oehmke)
“Since the algorithms learn from data, it’s not as easy to understand what they do as it would be if they were programmed by us, like traditional algorithms. But that’s the essence of machine learning: that it can go beyond our knowledge to discover new things. A phenomenon may be more complex than a human can understand, but not more complex than a computer can understand.

And in many cases we also don’t know what humans do: for example, we know how to drive a car, but we don’t know how to program a car to drive itself. But with machine learning the car can learn to drive by watching video of humans drive.” (*)

--- Pedro Domingos

(*) Source: On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos, ODBMS Industry Watch, June 18, 2018
Waymo (a subsidiary of Alphabet Inc) created a Recurrent Neural Network (RNN) for Driving.

They trained the neural network Imitating the “Good” and synthesizing the “Bad”.

„They trained the model with examples from the equivalent of about 60 days of expert driving data, while including training techniques such as past motion dropout to ensure that the network doesn’t simply continue to extrapolate from its past motion and actually responds correctly to the environment.“

(*) Source: Learning to Drive: Beyond Pure Imitation

“It's not difficult to feed the bad examples. That's what we do in our training, we feed it synthesized bad examples and add a training loss that tells the network not to emulate the bad behavior.

Real examples of bad behavior are difficult to intentionally obtain, and it is simpler and safer to synthetically create bad examples in simulation.”

--Abhijit Ogale

Source: Personal communication
“If the learning took place before the car was delivered to the customer, the car’s manufacturer would be liable, just as with any other machinery. The more interesting problem is if the car learned from its driver. Did the driver set a bad example, or did the car not learn properly?”

--Pedro Domingos

(*) Source: On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos, ODBMS Industry Watch, June 18, 2018
Causality – in other words, grasping not just patterns in data but why something happens. Why is that important, and why is it so hard?

“If you have a good causal model of the world you are dealing with, you can generalize even in unfamiliar situations. That’s crucial. We humans are able to project ourselves into situations that are very different from our day-to-day experience. Machines are not, because they don’t have these causal models.

We can hand-craft them, but that’s not enough. We need machines that can discover causal models. To some extent it’s never going to be perfect. We don’t have a perfect causal model of the reality; that’s why we make a lot of mistakes. But we are much better off at doing this than other animals.

Right now, we don’t really have good algorithms for this, but I think if enough people work at it and consider it important, we will make advances.”

-Yoshua Bengio

(™) Source MIT Technology Review
"Knowing why an expert driver behaved the way they did and what they were reacting to is critical to building a causal model of driving. For this reason, simply having a large number of expert demonstrations to imitate is not enough. Understanding the why makes it easier to know how to improve such a system, which is particularly important for safety-critical applications.” (*)

However, I do not believe that we know WHY and HOW we drive though…

Try for yourselves: Explain to another person how do you drive and why you react in certain situations they way you do….. And please let me know the result.

(*) Source: Learning to Drive: Beyond Pure Imitation
https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2
As a layperson looking at this particular field of ethical systems, I see some parallels between determining whether a system has intelligence and whether a system is making ethical decisions or not. In both cases, we are faced with a kind of Turing test scenario where we find it difficult to articulate what we mean by intelligence or ethics, and can only probe a system in a Turing test manner to determine that it is indistinguishable from a model human being.

The trouble with this approach though is that we are assuming that if the system passes the test, it shares the same or similar internal representations as the human tester, and it is likely that its intelligence or ethical behavior generalizes well to new situations. We do the same to assess whether another human is ethical or not.

This is a great difficulty, because we currently know that our artificial ML systems learn and generalize differently than humans do, so this kind of approach is unlikely to guarantee generally intelligent or ethical behavior.

I think the best we can currently do is to explicitly engineer/bound and rigorously test the system against a battery of diverse scenarios to check its decisions and reduce the likelihood of undesirable behavior.

The number of tests needs to be large and include long-tail scenarios because deep learning systems don't have as large a generalization horizon as human learning, as evidenced by their need of a mountain of training data. “

--- Abhijit Ogale

Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.
AI and the Rise of (Digital) Ecosystems

- The Rise of (Digital) Ecosystems paving the way to disruption.

- Different Countries, Different Approaches and Values. (China, the United States, Russia, Europe…)
  China has 2 completed ecosystems (Tencent, Alibaba)
  Russia is almost there.

- What does it have to do with us?

Source: Digital Hospitality, Metro AG
"Big Nudging"
He who has large amounts of data can manipulate people in subtle ways. But even benevolent decision-makers may do more wrong than right.

Spotlight on China: Is this what the Future of Society looks like?

How would *behavioural* and *social control* impact our lives? The concept of a Citizen Score, which is now being implemented in China, gives an idea.

“Citizens and businesses alike need to be able to trust the technology they interact with, and have effective safeguards protecting fundamental rights and freedoms.

In order to increase transparency and minimise the risk of bias, AI systems should be developed and deployed in a manner that allows humans to understand the basis of their actions.

Explainable AI is an essential factor in the process of strengthening people’s trust in such systems.” (*)

-- Roberto Viola
Director General of DG CONNECT (Directorate General of Communication Networks, Content and Technology) at the European Commission.

(*) Source On the Future of AI in Europe. Interview with Roberto Viola, ODBMS Industry Watch2018-10-09
EU High-Level Expert Group on AI presented their ethics guidelines for trustworthy artificial intelligence:

- (1) lawful - respecting all applicable laws and regulations
- (2) ethical - respecting ethical principles and values
- (3) robust - both from a technical perspective while taking into account its social environment
The guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy.

- **Human agency and oversight**: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

- **Technical Robustness and safety**: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

- **Privacy and data governance**: Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

- **Transparency**: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system’s capabilities and limitations.

- **Diversity, non-discrimination and fairness**: Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

- **Societal and environmental well-being**: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

- **Accountability**: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.
7. Accountability

- **Auditability:** Did you establish mechanisms that facilitate the system’s auditability, such as ensuring traceability and logging of the AI system’s processes and outcomes? Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

- **Minimising and reporting negative Impact:** Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected? Did you provide training and education to help developing accountability practices? Do these trainings also teach the potential legal framework applicable to the AI system? Did you consider establishing an ‘ethical AI review board’ or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas? Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives? Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

- **Documenting trade-offs:** Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them? How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

- **Ability to redress:** Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact? Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?
Accountable AI?

Do we need some sort of auditing tool?

The technology has to be able to “explain” itself, to explain how a data-driven algorithm came to the decision or recommendation that it did. Is it technically feasible?

This is current research work area: e.g

*Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi (Submitted on 14 Jan 2019)

How much *Transparency* is desired/ possible/ allowed….?

Do we wish “*Human in the loop*” for most of these kinds of decisions for the foreseeable future?
AI and The Paradox of Transparency

I do not mean cognitive biases…

I mean, if we really insist on AI Transparency, perhaps this would force us to reveal our real motives…

But, we do not always wish to make our motives visible to the outside world, e.g. we do not wish transparency….

But with no transparency, there is a lack of trust.
Who is responsible?

AI system designers and their managers do have ethical responsibilities. 

and

Other stakeholders (e.g. policy makers, politicians, opinion leaders, educators) do have ethical responsibilities.
What about the Citizens?

What is the implication for them of an AI Ethical Inspection?

Shall we involve them as well? How?
Why do we need an AI Ethical Due Diligence?

- **Forensic investigation** → Minimize Risks

- ..... 

- **Holistic investigation** → Being Ethical

**ethical**

/ˈɛθɪk(ə)l/

*Adjective*

relating to moral principles
How to Use an AI Ethical Due Diligence

Two different ways to use it:

As part of an Ethics by Design process,

or if the AI has already been designed, it can be used to do an

*Ethical sanity check*, so that a certain AI Ethical standard of care is achieved.

It can be used by a variety of AI stakeholders.

Contribute to closing the gap between “*principles*” (the “*what*” of AI ethics) and “*practices*” (the “*how*”).
Ethics "embedded" into the core of the AI design. Not reacting to it....

Kind of "Ethics inside".

Nice in theory, but difficult to enforce it in practice at scale (different ecosystems, different value systems, business goals,...)
How to do an AI Ethical *sanity* check?
An holistic AI Ethical sanity check

This is going to happen soon, especially the Forensic investigation \( \rightarrow \) Minimizing Risks

We want to do more than that, and create and test with real use cases an holistic AI Ethics Inspection Process, which goes beyond minimizing risks and include assessing how ethical is the overall context in which AI is designed/produced/used.
An Holistic Ethical Inspection
AI, People, Ecosystems,...
Z-inspection:
An Ethical AI Inspection
Using the 7 key requirements set by the EU experts that AI systems should meet in order to be deemed trustworthy
What do we wish to investigate?

AI is not in isolation. It is part of:
- One or more (digital) ecosystems
- Processes, Products, Services, etc.

It is related to People, Data, Values, Laws, Context

AI is made up of various components (e.g. deep neural network architectures: neural networks building blocks)
Focus of the AI Ethics Inspection

- Legal
- Ethical
- Technical

**Note1:** Illegal and unethical are not the same thing.

**Note2:** Legal and Ethics depend on the context.
Connecting AI, Ecosystems, Ethical Values and Democracy

Is this part of an AI Ethical Inspection?
Z-Inspection Process

1. Define an holistic Methodology/ Extend Existing Validation Frameworks and Practices to assess and mitigate risks and undesired “un-ethical side effects”.

   - Use/ Develop new Tools, Use/ Extend existing Toolkits, Use/
   - Use/ Define ML Fairness Metrics,
   - Define Ethics AI benchmarks
   - Define Scenarios (Data/ Process/ People / Ecosystems)

2. Create a Team of inspectors

3. Involve relevant Stakeholders

4. Apply Methodology/ Tools/ Metrics to Real Use Cases

5. Manage risks/ Remedies (when possible)

6. Feedback: Learn from the experience

7. Iterate: Refine Methodology / Develop Tools…

Other?
Ethical AI "Macro"-Investigation

(Digital) ECOSYSTEM X

X,Y,Z = US, Europe, China, Russia, others…
Ethical AI “Micro”-Investigation

VALUES

VALUES CHECK

“Good”

“Bad”

Context Culture People/Company Values

People + Algorithms + Data

AI

Feedback

“Good”

???
They have used 30 million of real-world “expert” driving examples.

Q. How did you define an “expert” in your work? In the blog they speak of a “good driver”. Who is a good driver?

In the paper they write that “this is a difficult robotics challenge (i.e. predict, plan) that humans solve well”.

Q. How do you define if a human solve this well? This is particularly true when logically unexpected and unpredictable things happens on the road (a fire, a hearth quake, a bridge collapses, etc.)

They design a RNN to output a trajectory which consists of ten future points.

Q. Why ten? Any particular rational for this?

They define an “imitation dropout” as composed of imitation losses plus environment losses.

Q. How is the learning (and accuracy) affected if you change the dropout strategy?
Micro-Validation: Is it sufficient?

AI (Ethically Checked)!

• Passed the test
Micro-validation does not imply Macro-validation

AI Ethically Checked!
An idea is to start with a Micro-Investigation and then if needed progressively extend it in an incremental fashion to include a Macro-Investigation.

The output of this investigation is a degree of confidence that the AI analyzed - taking into account the context (e.g. ecosystems), people, data and processes - is ethical with respect to a scale of confidence. E.g. Score=9 Very strong, Score=6 Low but sufficient. Below 6 = risky.
Z-inspection Trade-off concept

Trade-off (what can go wrong?)

*Using AI vs Non using AI for the specific use case*

The Decision is Domain specific and depends on the Context, Use case, and the Ethical Score obtained after inspection (6-9, below 6).

*Responsibility, Consequences, Ethical?*
Perhaps we can “certify” AIs by the number of testing with synthetics data sets and extreme scenario they went through before allowing AIs to drive a car (similar to what happens to airplane pilots).

Somebody would need to define when good is enough. And this may be tricky…

More feedback I have received, and resources here:
How often should be inspected?

- Need to define a set of checkpoints that need to be monitored over time
- For minimal inspection and full inspection.
- How to cope with Changes over time (Ecosystems, Ethical values, technological progress, research results, etc.)

Note: This may result in a ethical inspection of the entire context in which AI is deployed. Could raise issues and resistance..
Are there any tools around?

Ethics & Algorithms Toolkit
A risk management framework for governments (and other people too!)
GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC have collaborated on a practical toolkit for cities to use to help them understand the implications of using an algorithm, clearly articulate the potential risks, and identify ways to mitigate them.

https://ethicstoolkit.ai/
“In Part 1 of the toolkit, there are six major steps (or questions) to help you and your stakeholders characterize an algorithm. Many of these steps have multiple components, but also include clear instructions on how to summarize those stages in order to complete the step.”

Source: https://ethicstoolkit.ai/
Part 1: Assess Algorithm Risk (cont.)

1. **Impact** of algorithm wrt people and property
2. **Appropriate use** if the data and algorithm are used for the purpose anticipated and perception of use
3. **Accountability** role of people in the use of algorithm and if automated decisions can be explained
4. **Bias** explores the undelining influence of the data and people who help build the algorithm

Source: https://ethicstoolkit.ai/
“Although it’s helpful to know how concerned you should be about various aspects of your algorithm, that’s really only half the battle. Although there may be a few cases where the risks are too severe to proceed, there are often ways to mitigate them. Using Part 2 of the toolkit, you identify specific techniques to help address the considerations you identified in Part 1.”

Source: https://ethicstoolkit.ai/
For Step 3.3 “accountability”...

- If you selected **low** or **medium**, use automated testing tools to periodically evaluate algorithm performance (mitigation 7), ensure there is a human adjudication mechanism (mitigation 8), and require human intervention before executing each algorithmic decision (mitigation 9).

- If you selected **high**, ensure human adjudication mechanism results feed into algorithm tuning (mitigation 8), ensure the relevant inputs and machine state(s) are captured in perpetuity for each decision (mitigation 10), and evaluate human-intervened decisions periodically (mitigation 11).

- **Mitigation 7.** Automating testing tools (i.e. confusion matrices when evaluating classification models) to evaluate an algorithm’s performance can be a way to integrate systematic checks into the lifecycle of an algorithm. If the aforementioned classification model is falsely classifying 70% of cases, the automated testing tool can be programmed to produce “STOP” in red letters.

- **Mitigation 8.** A human adjudication mechanism, being a process through which a person can introduce his or her own discernment, can be a great addition to a project involving an algorithm or algorithms. Ensuring that this mechanism can then feed into the tuning of an algorithm can be a great addition to the project.

- **Mitigation 11.** Evaluate human-intervened decisions periodically to control for unintended rater bias.
Ethics Toolkit ++

**Ethics & Algorithms Toolkit**

<table>
<thead>
<tr>
<th>Assess third party methodology risk</th>
<th>Understand and assess impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assess appropriate data use risk</td>
<td>Determined automation</td>
</tr>
<tr>
<td>Determine explainability</td>
<td>Determine auditability</td>
</tr>
<tr>
<td>Bias in data</td>
<td></td>
</tr>
<tr>
<td>Bias in developing and operating teams</td>
<td></td>
</tr>
</tbody>
</table>

**In Common**

<table>
<thead>
<tr>
<th>ETHICS GUIDELINES FOR TRUSTWORTHY AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI and human interaction</td>
</tr>
<tr>
<td>Attacks and misuse</td>
</tr>
<tr>
<td>Reliance - technical availability, accuracy</td>
</tr>
<tr>
<td>Privacy of the used data</td>
</tr>
<tr>
<td>Data governance</td>
</tr>
<tr>
<td>Accessibility for users</td>
</tr>
<tr>
<td>Feedback of stakeholders</td>
</tr>
</tbody>
</table>

Part 1

Part 2

Part 2 Ethics Toolkit ++
Additional Toolkits

- AI Fairness 360 (IBM)
- What-if Tool (Google)
- Aequitas (Univ. Chicago)
- Lime (Univ. Washington)
- DotEveryone Consequence Scanning Event
- …
We plan to apply Z-Inspection
Use Case: CARDISIO

Company description: http://cardis.io/

Motivation for AI- supposed benefits:

“The first highly accurate and non-invasive test to determine a risk factor for coronary heart disease.

Easy to use. Anytime. Anywhere.”
Why Being Ethical: Incentives?

“Evaluation and creation of pro-ethical business models and incentive structures that balance the costs and rewards of investing in ethical AI across society”

Military Applications for Artificial Intelligence

Drones That Kill on Their Own: Will Artificial Intelligence Reach the Battlefield?

Just war theory (Latin: jus bellum justum) is a doctrine, also referred to as a tradition, of military ethics studied by military leaders, theologians, ethicists and policy makers. The purpose of the doctrine is to ensure war is morally justifiable through a series of criteria, all of which must be met for a war to be considered just. The criteria are split into two groups: "right to go to war" (jus ad bellum) and "right conduct in war" (jus in bello). The first concerns the morality of going to war, and the second the moral conduct within war. Recently there have been calls for the inclusion of a third category of just war theory — jus post bellum — dealing with the morality of post-war settlement and reconstruction. (source Wikipedia)
AI and Science Fiction

2001: A Space Odyssey

is a 1968 science fiction novel by British writer Arthur C. Clarke. It was developed concurrently with Stanley Kubrick's film version and published after the release of the film.

Are we talking about Science Fiction here or...

“The ideal of General AI is that the system would possess the cognitive abilities and general experiential understanding of its environments that we humans possess, coupled with the ability to process this data at much greater speeds than mortals. It follows that the system would then become exponentially greater than humans in the areas of knowledge, cognitive ability and processing speed – giving rise to a very interesting species-defining moment in which the human species are surpassed by this (now very) strong AI entity.”

Source: https://hackernoon.com/general-vs-narrow-ai-3d0d02ef3e28

and this poses severe Ethical concerns....

I also believe that AI initiative such as Neuralink: https://www.neuralink.com/ poses serious Ethical issues...

„Creating a neural lace is the thing that really matters for humanity to achieve symbiosis with machines“ -- Elon Musk (*)

"I think ethical software development for AI is not fundamentally different from ethical software development in general. The interesting new question is: when AIs learn by themselves, how do we keep them from going astray?

Fixed rules of ethics, like Asimov’s three laws of robotics, are too rigid and fail easily. (That’s what his robot stories were about.) But if we just let machines learn ethics by observing and emulating us, they will learn to do lots of unethical things.

So maybe AI will force us to confront what we really mean by ethics before we can decide how we want AIs to be ethical." (*)

--Pedro Domingos (Professor at University of Washington)
To me, it seems that AI is in the category of disruptive technologies, such as Nuclear technology, only more pervasive....

ECOSYSTEMS XYZ: AI, “Embedded” AI in products and services is changing our Life(Style).
And they said, Go to, let us build us a city and a tower, whose top may reach unto heaven;

— *Genesis 11:4*