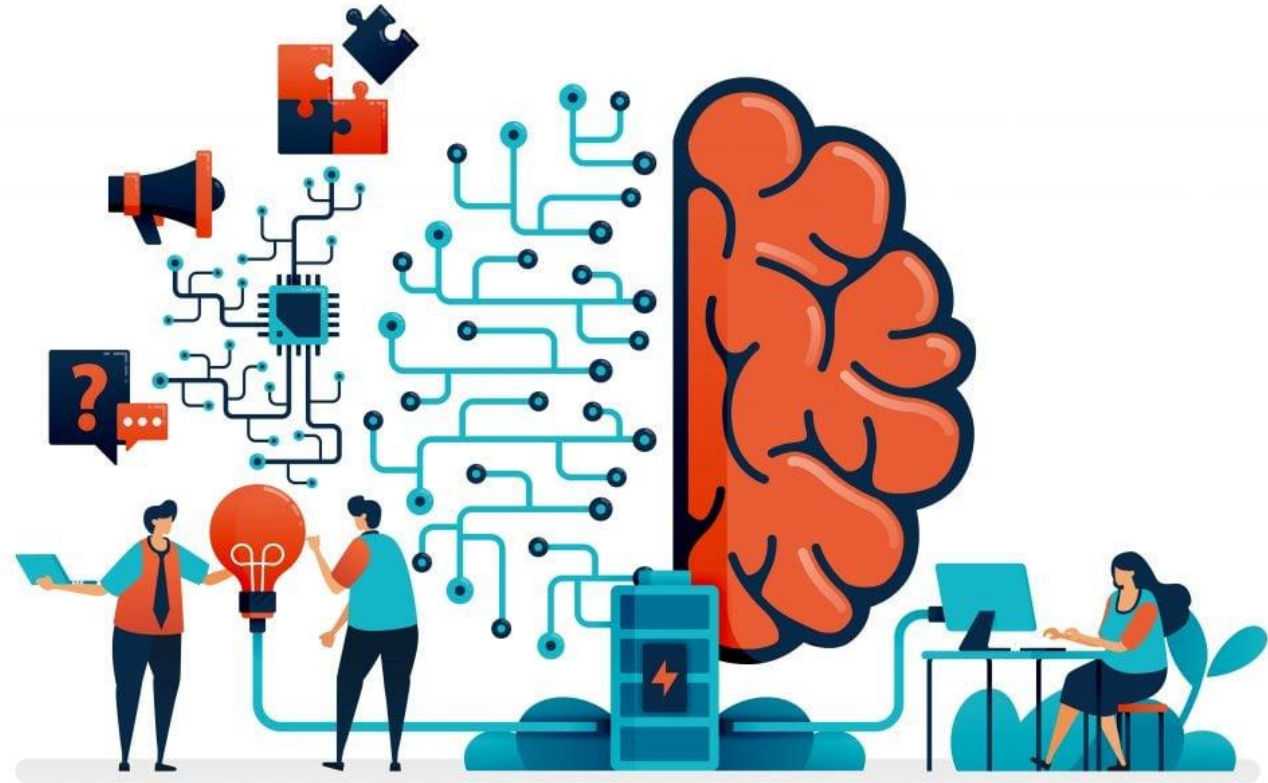


Towards an Ethical Artificial Intelligence

Consolidating Transparency, Fairness and Accountability in Machine Learning Models





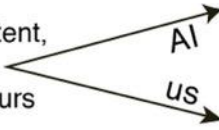








Alberto Fernández – Instituto DASCI . Universidad de Granada



Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter (2015)

Fdo. Stephen Hawking, Elon Musk, DeepMind, Vicarious, Peter Norvig...

The potential benefits (of AI) are enormous, since everything civilization offers is the product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified with the tools AI can provide, but the eradication of disease and poverty are not unsuspected. Because of AI's great potential, it is important to investigate how to harness its benefits while avoiding potential pitfalls.

<p>Myth: Superintelligence by 2100 is inevitable</p> <p>Myth: Superintelligence by 2100 is impossible</p>	<table border="1"> <tr><th>Mon</th><th>Tue</th><th>Wed</th><th>Thu</th><th>Fri</th><th>Sat</th><th>Sun</th></tr> <tr><td></td><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr> <tr><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td></tr> <tr><td>12</td><td>13</td><td>14</td><td>15</td><td>16</td><td>17</td><td>18</td></tr> <tr><td>19</td><td>20</td><td>✓ 21</td><td>22</td><td>23</td><td>24</td><td>25</td></tr> <tr><td>26</td><td>27</td><td>28</td><td>29</td><td>30</td><td></td><td></td></tr> </table>	Mon	Tue	Wed	Thu	Fri	Sat	Sun				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	✓ 21	22	23	24	25	26	27	28	29	30			<p>Fact: It may happen in decades, centuries or never: AI experts disagree & we simply don't know</p>	
Mon	Tue	Wed	Thu	Fri	Sat	Sun																																							
			1	2	3	4																																							
5	6	7	8	9	10	11																																							
12	13	14	15	16	17	18																																							
19	20	✓ 21	22	23	24	25																																							
26	27	28	29	30																																									
<p>Myth: Only Luddites worry about AI</p>		<p>Fact: Many top AI researchers are concerned</p>																																											
<p>Mythical worry: AI turning evil</p> <p>Mythical worry: AI turning conscious</p>		<p>Actual worry: AI turning competent, with goals misaligned with ours</p>																																											
<p>Myth: Robots are the main concern</p>		<p>Fact: Misaligned intelligence is the main concern: it needs no body, only an internet connection</p>																																											
<p>Myth: AI can't control humans</p>		<p>Fact: Intelligence enables control: we control tigers by being smarter</p>																																											
<p>Myth: Machines can't have goals</p>		<p>Fact: A heat-seeking missile has a goal</p>																																											
<p>Mythical worry: Superintelligence is just years away</p>	<p>PANIC!</p> 	<p>Actual worry: It's at least decades away, but it may take that long to make it safe</p>	<p>PLAN AHEAD!</p> 																																										

INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

Artificial
Intelligence
Governance



5

Going deeper in
the study areas
on AI ethics



6

Summary: AI
implications, *AI
for Good*,
suggestions

INDEX



1

**Natural
Intelligence vs.
Artificial
Intelligence**



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

Artificial
Intelligence
Governance



5

Going deeper in
the study areas
on AI ethics



6

Summary: AI
implications, *AI
for Good*,
suggestions



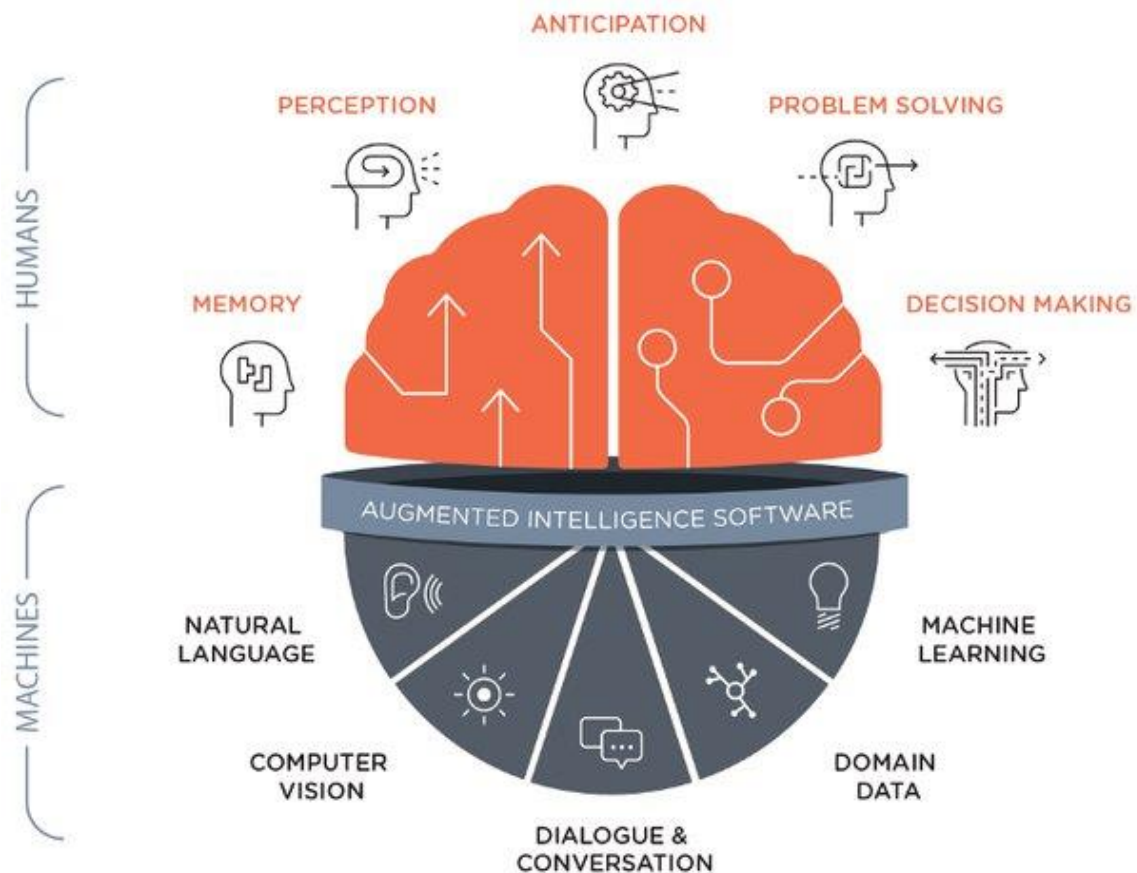
What is intelligence?

- Faculty of the mind that allows **learning**, understanding, reasoning, making **decisions** and forming a specific idea of **reality**.
- Ability to acquire **knowledge**, to think and reason effectively, and to manage in the environment in an **adaptive** way.
- Ability to **solve** problems or produce valuable goods.





Human vs. Artificial Intelligence



- Computational / robotic systems designed to perform (**mimic**) activities that can be performed by humans:
 - Reasoning, Decision, Acting.
- Weak AI (specific) vs. Strong AI (general):
 - Application specific vs.
 - Consciousness, Wisdom, Self-awareness: Singularity?



What concepts do we associate with AI?



Algorithms

- “Recipe” adaptive to context



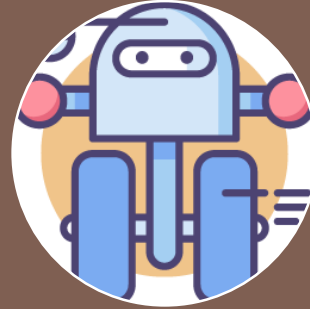
Data

- Data “curation” / “governance”
- Sustainability



Machine & Deep Learning

- High level feature extraction and pattern recognition



Robotics



Decision making



Human interactions

- A better world? The end of days?



Current implications of AI



Digital transformation

Imply new learnign skills

Does it revert on digital Illiteracy?

4.0 Revolution

Unprecedented technological developments

Widens the social gap?

Economic Impact

Increases growing rate

Job loses?

Substitutes identification and decision making

Speed and robustness

Do we have full confidence on systems?

INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

Artificial
Intelligence
Governance



5

Going deeper in
the study areas
on AI ethics



6

Summary: AI
implications, *AI*
for Good,
suggestions



What is ethics? Philosophical point of view

- Associated questions: "What is a good deed?", "What is the value of human life?", "What is justice?" or "What is a good life?".
- The terms "moral" and "ethical" are frequently used in ethical discussions.
 - **Moral:** patterns of behavior, customs and convention of cultures, groups or individuals.
 - **Ethical:** evaluating actions and behaviors from a systematic and scholarly perspective.
- Multitude of currents, sometimes contradictory: Utilitarian, Kantian, Aristotelian.
- They do not provide methodologies for conflict resolution





What is the ethics of Artificial Intelligence?

A set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development, deployment, and use of AI technologies.

FATE

“Fairness”, “Accountability”,
“Transparency”, and “Ethics”

D. Leslie, *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, 2019



What about our own ethics?

- The trolley dilemma.
- You are invited, without further details, to make a "safe" investment x10 from 500\$.
- A friend confesses to a crime. The next day the news announces a murder.
- When making a purchase you are given too much change.
- You receive a package delivered to the wrong address with a favorite item you can't buy yourself.
- A neighbor doesn't take good care of your pet.



Example of "ethical" car design

Utilitarian car

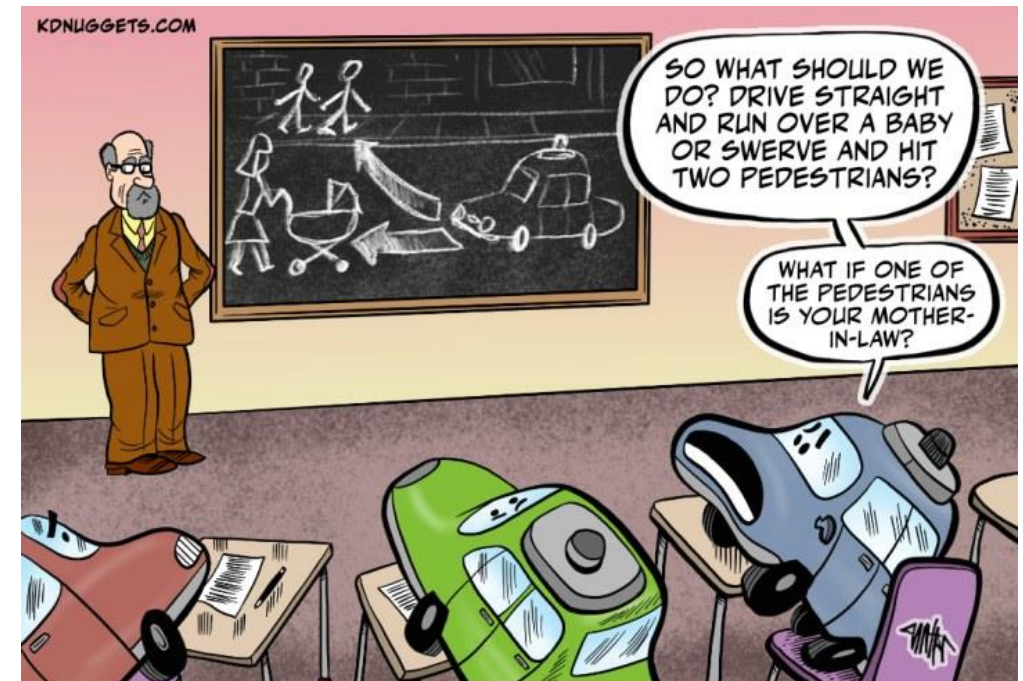
- What's best for the majority; results matter
- Maximizing lives

Kantian car

- Do no harm
- Do no explicit action if that action causes harm

Aristotelian car

- Pure motives; motives matter
- Harm as little as possible; avoid the least advantaged (pedestrians?)

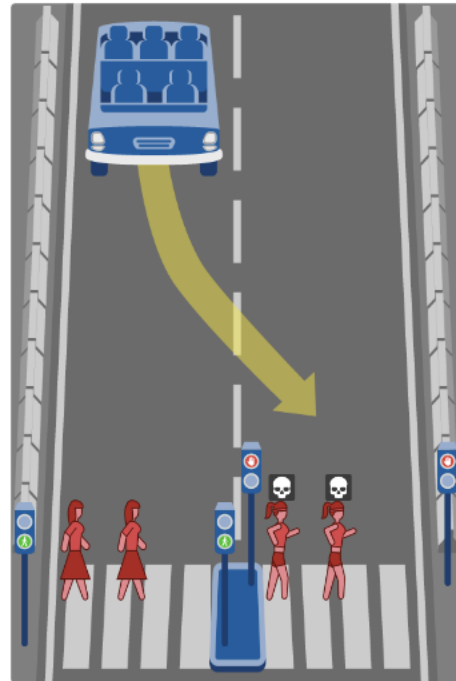
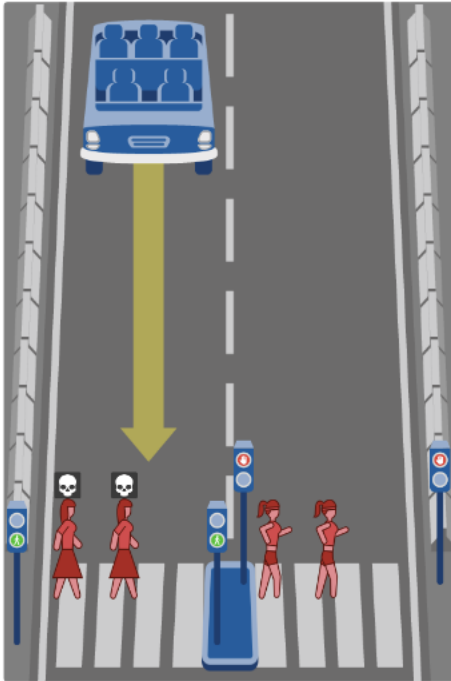




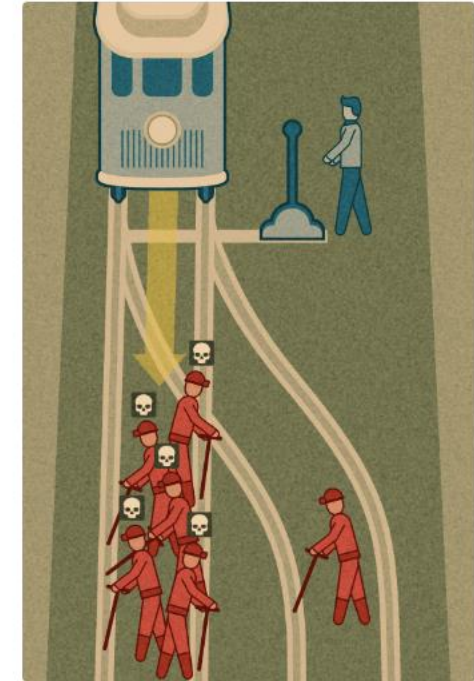
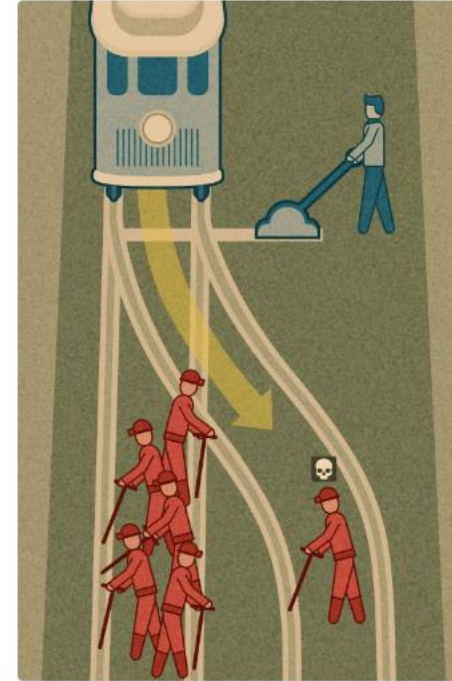
Practical Exercise: Moral Machine



What should the self-driving car do?



What should the man in blue do?





Sensitive use cases and "dystopias".

Denial of derivative services:

- Financial, housing, insurance, education, employment, health care services, etc.

Risk of harm:

- Physical, emotional or psychological to an individual (health care settings, etc.).

Infringement of human rights:

- Significant restriction of personal freedom, freedom of opinion or expression, freedom of assembly or association, privacy, etc. (facial recognition in police surveillance).



Current ethical dilemmas in AI

- Web Search Engines
 - Best leaders - gender?
 - Schoolgirl vs. Schoolgirl
- Classic dilemma of the accident
- Designation of authorship in artistic works, rights, integrity.
- Digitization of justice
 - Transparency / Trustworthiness
 - Privacy / Risks and rights...





Examples of "bad" use of AI

<p>VERNON PRATER</p> <p>Prior Offenses 2 armed robberies, 1 attempted armed robbery</p> <p>Subsequent Offenses 1 grand theft</p> <p>LOW RISK 3</p>	<p>BRISHA BORDEN</p> <p>Prior Offenses 4 juvenile misdemeanors</p> <p>Subsequent Offenses None</p> <p>HIGH RISK 8</p>
---	--

COMPAS: Recidivism convicts

J. Angwin, J. Larson, L. Kirchner, y S. Mattu, «Machine Bias», ProPublica, May 2016.

Dissecting racial bias in an algorithm used to manage the health of populations

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Health risk treatments

S. Vartan “Racial Bias Found in a Major Health Care Risk Algorithm”, October 2019.

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

Amazon Hiring

J. Dastin, «Amazon scraps secret AI recruiting tool that showed bias against women», Reuters, oct. 10, 2018.

INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

**Principles of
Responsible
Artificial
Intelligence**



4

Artificial
Intelligence
Governance



5

Going deeper in
the study areas
on AI ethics



6

Summary: AI
implications, *AI
for Good*,
suggestions



Why is Ethics in AI important?

Opacity of the models



Difficulty in evaluating the quality of the models.



Low confidence

Lack of morals and ethics: discrimination and biases

Hardly justifiable use

Invasion of privacy

High impact on subjects

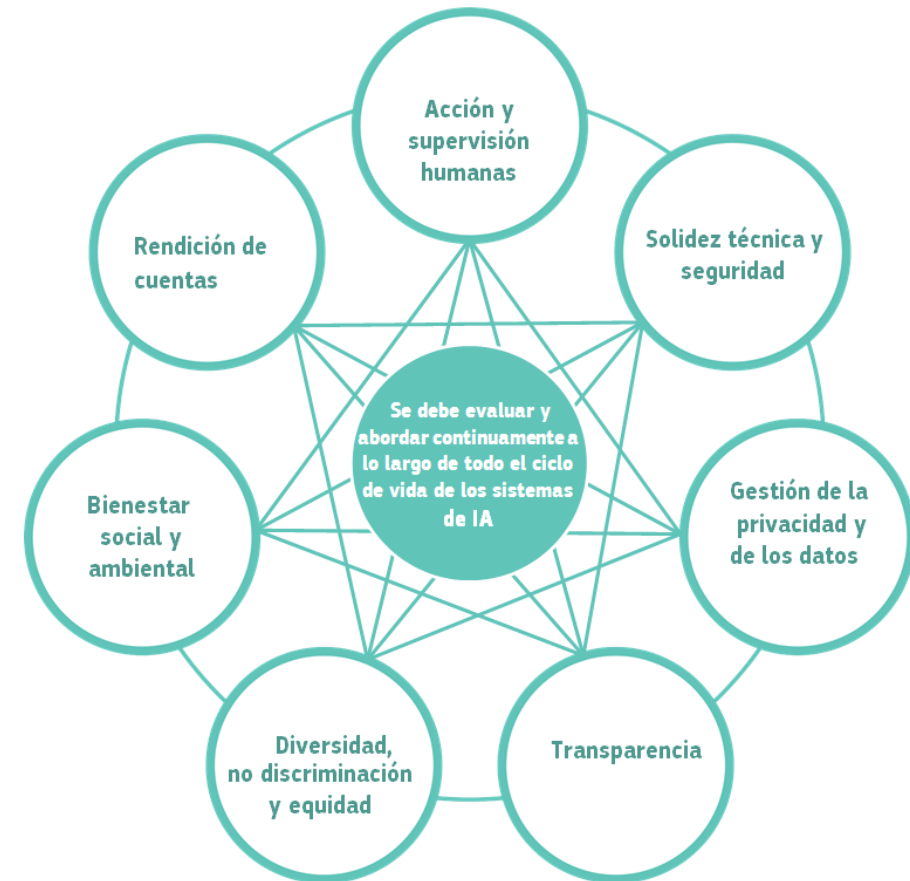
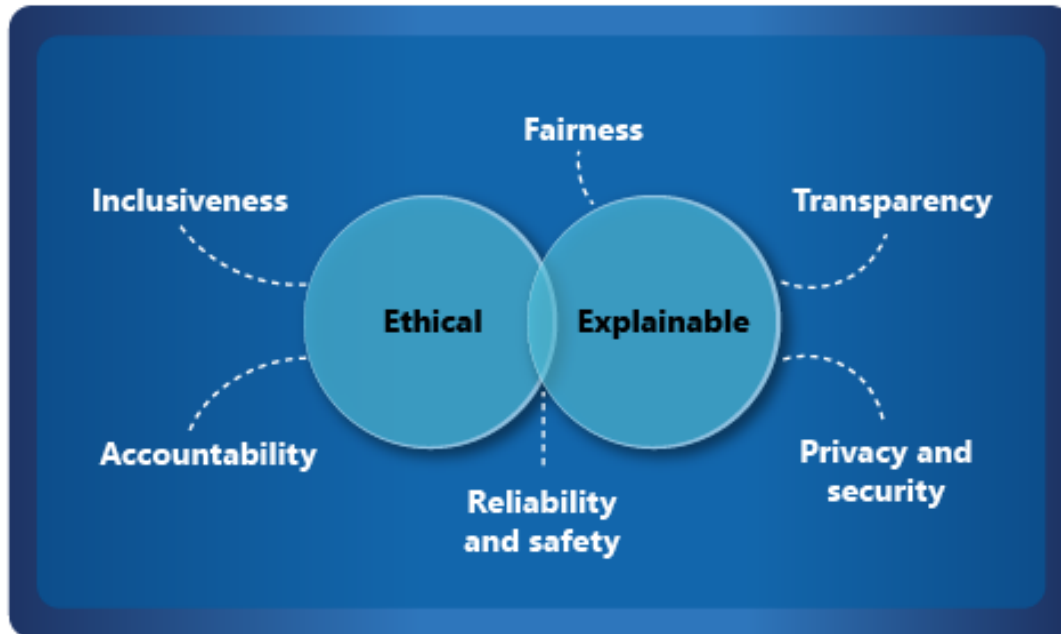
D. Leslie, *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute, 2019



Principles of Responsible AI



The principles of responsible AI



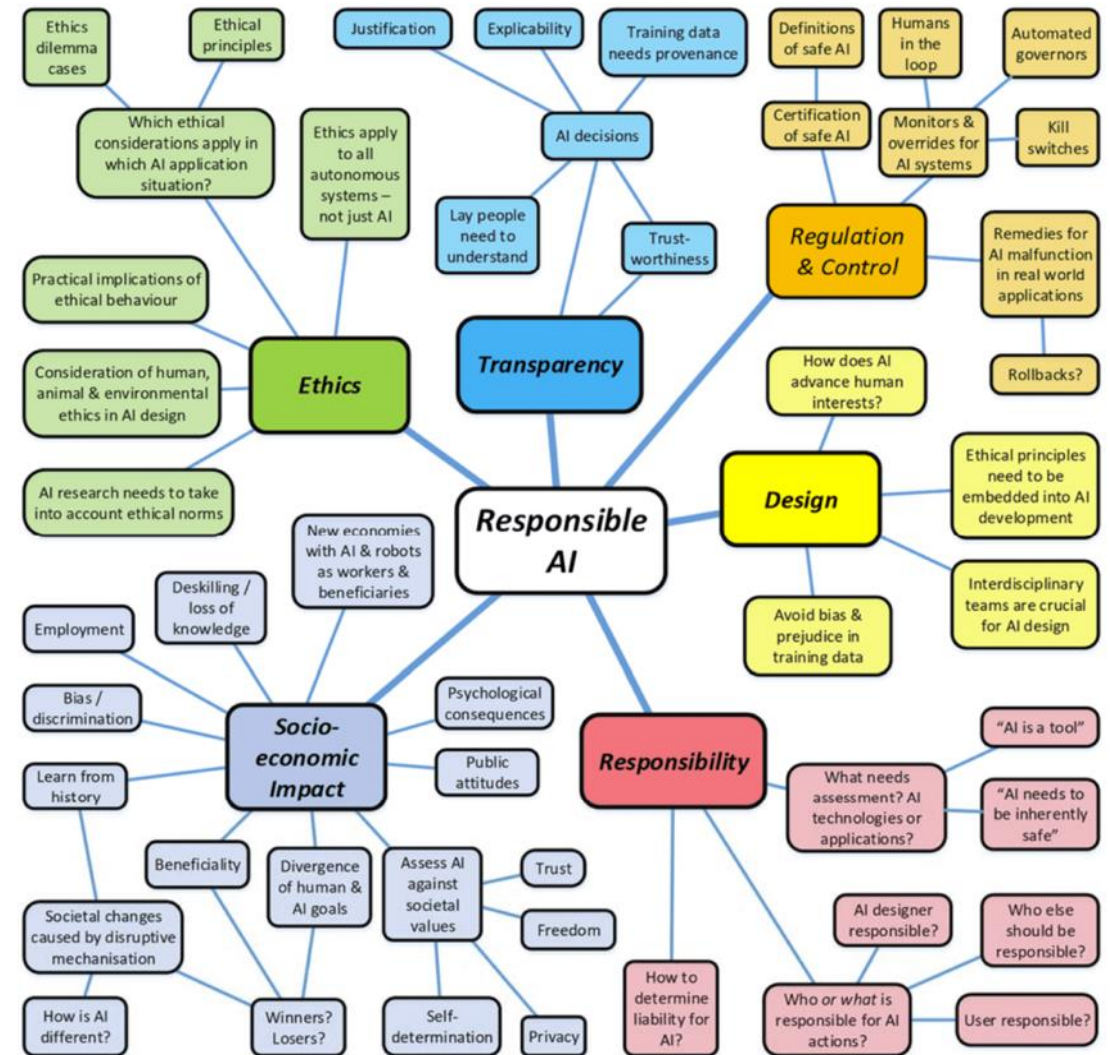


Principles of Responsible AI (2)

Explainable AI (XAI):	The ability to explain a model after it has been developed.
Interpretable Machine Learning:	Transparent model architectures and increase in how intuitive and understandable ML models can be
Ethical AI:	Sociological fairness in machine learning predictions (i.e., whether a category of people is being weighted unequally)
Safe AI:	Debugging and deployment of ML models with similar countermeasures against insider and cyber threats to what would be seen in traditional software
Human-centric AI:	User interaction/intervention/monitoring with AI and ML systems
Compliance:	Ensuring that your AI systems comply with relevant regulatory requirements , whether with GDPR, GDPR, GDPR, FCRA, or other regulations.

Principles of Responsible AI (and 3)

- **Ethics:** implications for AI
- **Transparency:** justifying and explaining AI decisions and actions
- **Regulation and control:** legislation and behavioral oversight
- **Socio-economic impact:** how are they affected by AI?
- **Design:** technical considerations
- **Accountability:** moral and legal responsibility.



INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

**Artificial
Intelligence
Governance**



5

Going deeper in
the study areas
on AI ethics



6

Summary: AI
implications, *AI
for Good*,
suggestions



GPAI /

THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

– Our values

Members and participants of GPAI are brought together first and foremost by a shared commitment to the values expressed in the [OECD Recommendation on Artificial Intelligence](#). All GPAI activities are intended to foster responsible development of AI grounded in these principles of **human rights, inclusion, diversity, innovation** and **economic growth**.

Principles for responsible stewardship of trustworthy AI

- › Inclusive growth, sustainable development and well-being
- › Human-centred values and fairness
- › Transparency and explainability
- › Robustness, security and safety
- › Accountability

National policies and international cooperation for trustworthy AI

- › Investing in AI research and development
- › Fostering a digital ecosystem for AI
- › Shaping an enabling policy environment for AI
- › Building human capacity and preparing for labour market transformation
- › International cooperation for trustworthy AI

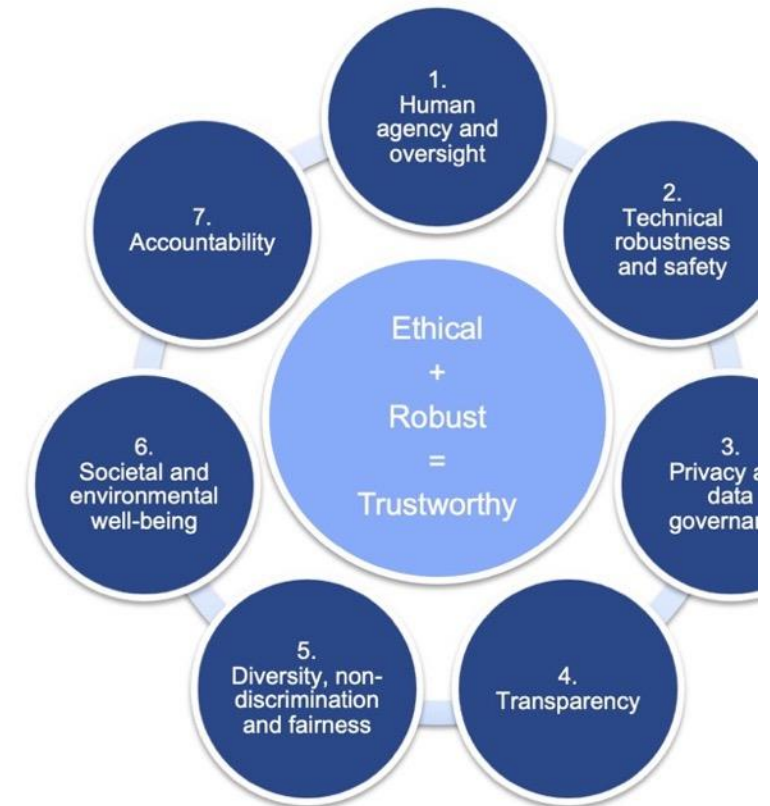


European regulations

- Regulations in occupational areas: architecture, military, medicine,
- All have a dependence on society
- Codes of conduct on mission, values and principles, linking them to rules and regulations.
- Compliance necessary to mitigate potential risks.
- EC Ethical Guideline: 7 requirements



Seven key requirements





AI-Act: Promoting safe AI that respects fundamental rights

Definition of AI system

- Distinguish AI from simpler software systems: machine learning approaches and logic and knowledge-based approaches.

Prohibited AI practices

- Avoid AI for social scoring.
- Exploiting vulnerabilities of groups due to their social or economic status.
- Use of biometrics only for law enforcement uses.

General purpose AI systems

- Provide for the use of AI with different application for use in situations of potential risk.



AI-Act: Identification of high-risk AI systems. <https://artificialintelligenceact.eu/>

Implement a trusted ecosystem on legal framework on the use of AI-based solutions, and encourage companies to develop them.

Enforce the goal of developing a trusted AI ecosystem.

Risk-based approach whereby legal intervention is tailored to mitigate the level of risk: e.g., biases, errors and opacity that may adversely affect a number of fundamental rights.



AI-Act: Identification of high-risk AI systems.

<https://artificialintelligenceact.eu/>

Biometric identification and categorization of individuals,

Management and operation of essential infrastructures,

Education and vocational training,

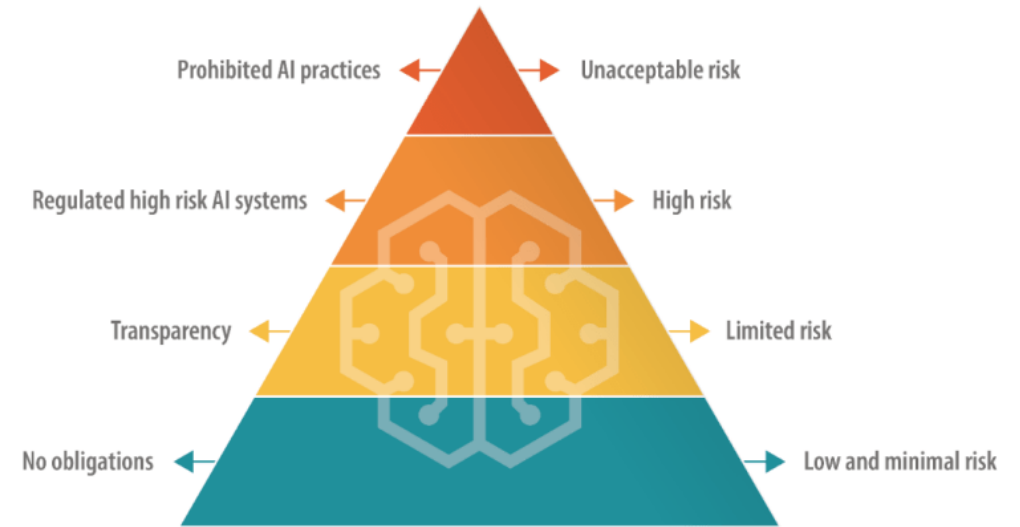
Employment, management of workers and access to self-employment,

Access to and enjoyment of essential public and private services and their benefits,

Law enforcement issues,

Migration, asylum and border control management, or

Administration of justice and democratic processes



Data source: [European Commission](https://europeancommission.eu).



Pilot project: AI Sandbox (Spain)

Creation (authorities and development companies) of a "good practice guide" and implementation guidelines:

- Obligations to be fulfilled and how to implement them.
- Method for control and monitoring

Increased legal certainty and transfer of know-how on compliance.

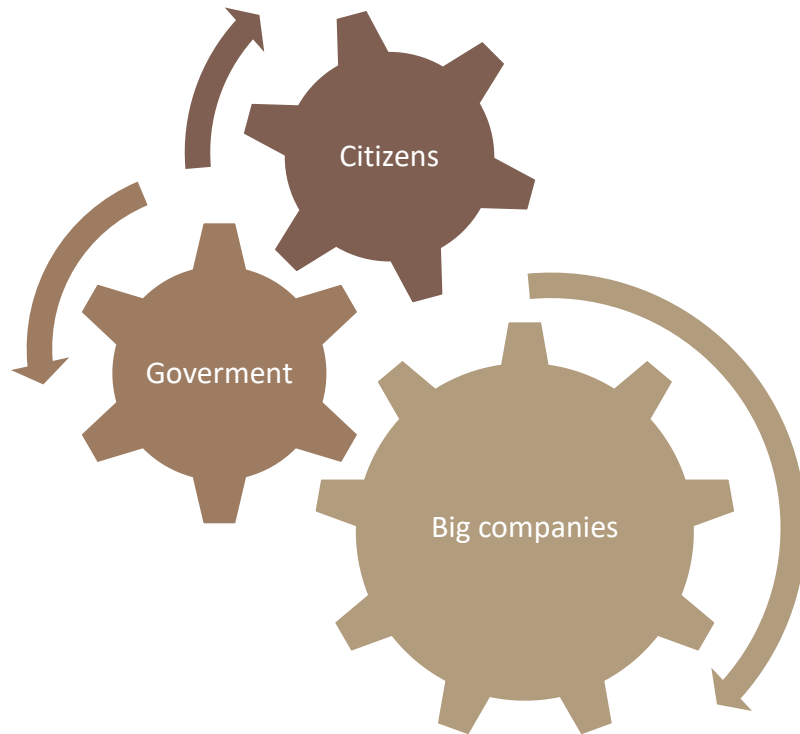
Basis for future European regulation (AI Act) in two years' time; associated with AESIA.





Governance: Europe vs. USA vs. China

Who has the data?



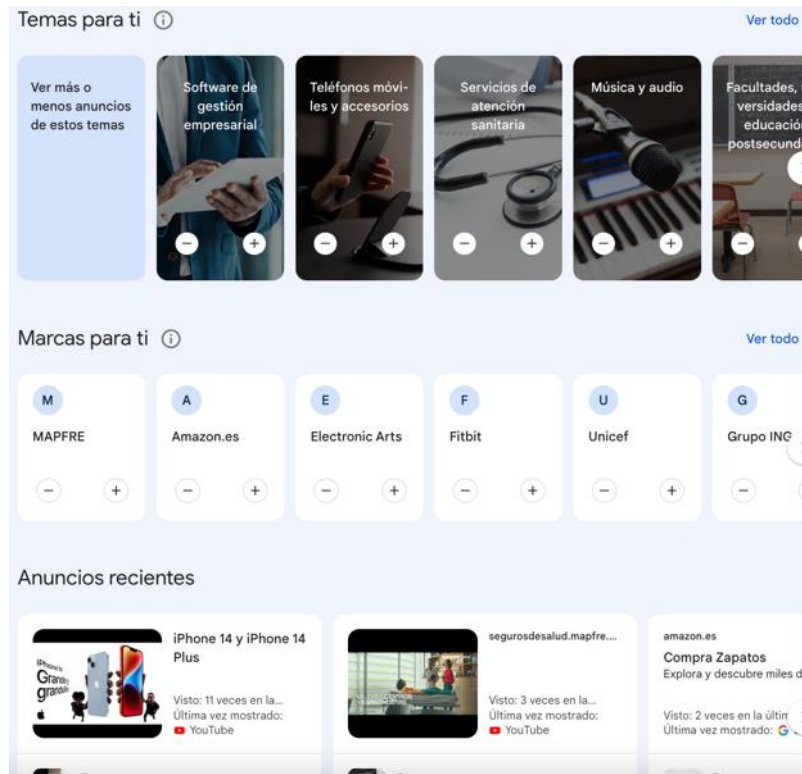
Data Brokers

- If you don't pay for a product, you are the product.
- *"Extensive and unregulated ecosystem."*
 - Enables companies to use personal data to target consumers
 - More than "targeted marketing."
- Intermediaries of "surveillance capitalism."
- They know more about you than you think, and do a lot more with it than you'd like,
- Is our data anonymous?
- Do we need these profiling services?

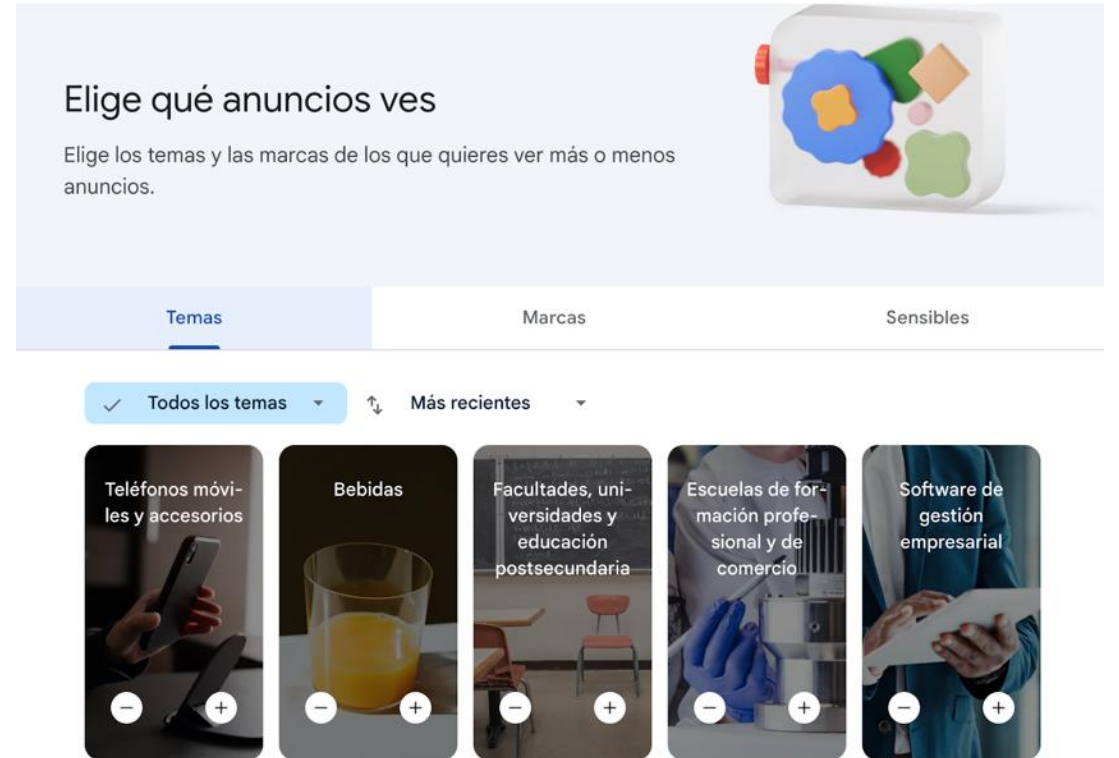


Example of profiling

Information on Google Ads



Customization (if known)



INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

Artificial
Intelligence
Governance



5

**Going deeper in
the study areas
on AI ethics**

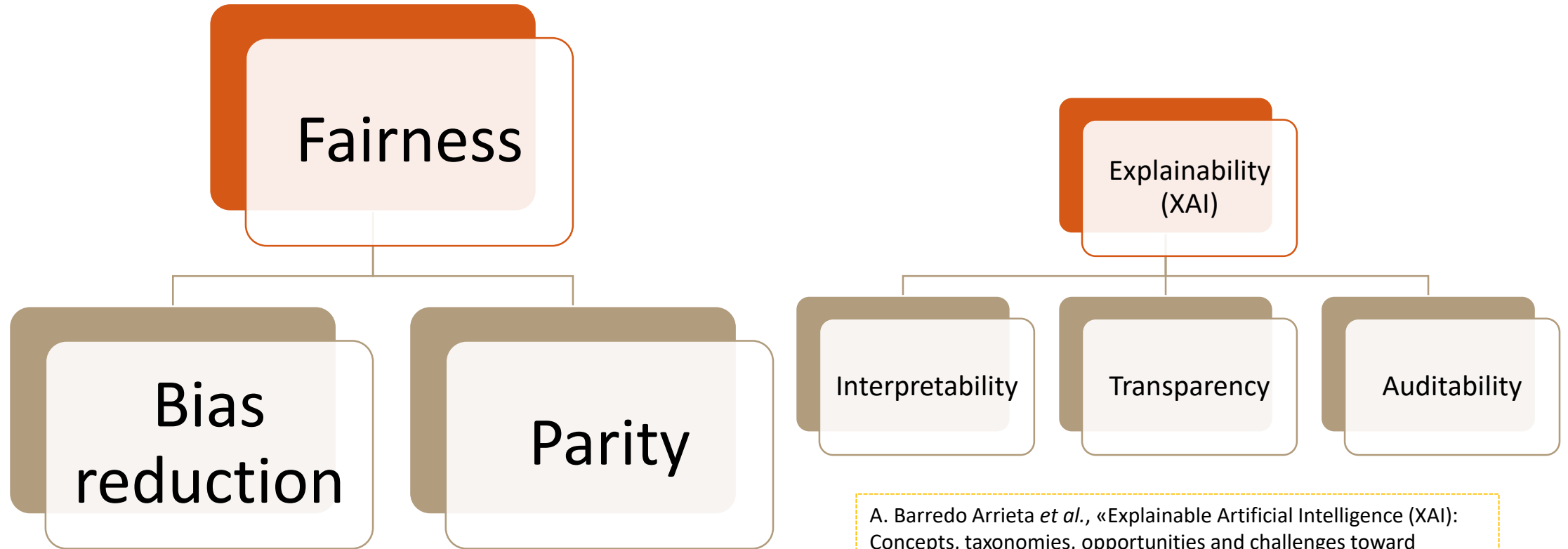


6

Summary: AI
implications, *AI
for Good*,
suggestions

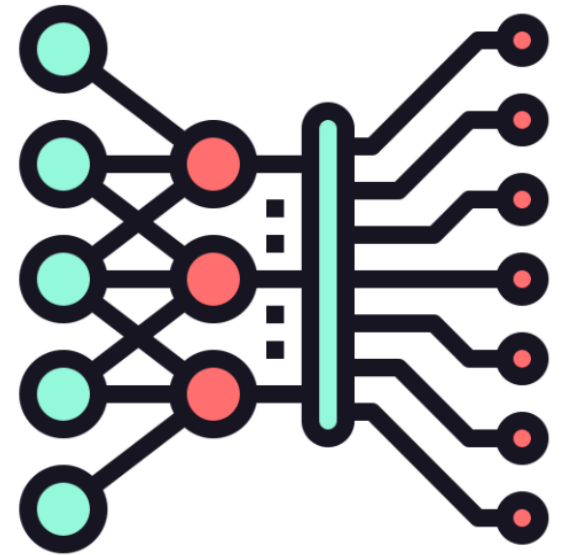
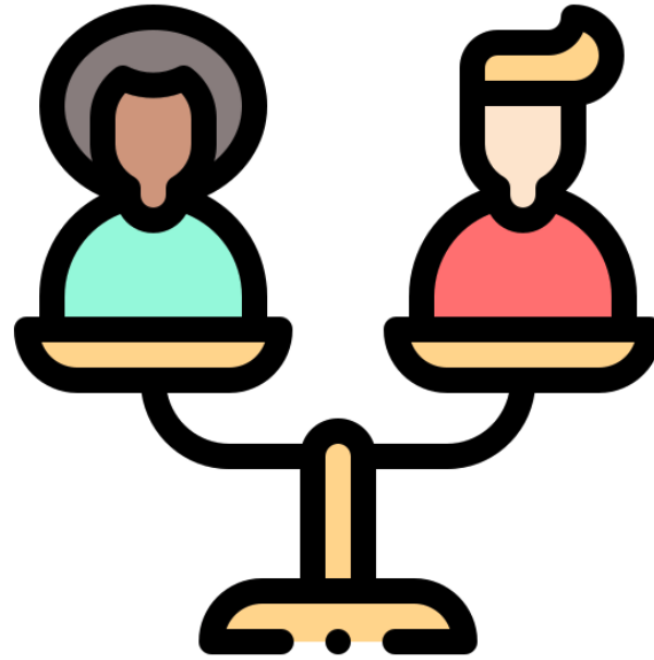


Branches of study of Ethics in AI



A. Barredo Arrieta *et al.*, «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI», *Inf. Fusion*, vol. 58, pp. 82-115, jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

FAIRNESS





Impartiality / fairness in IA

Sameness

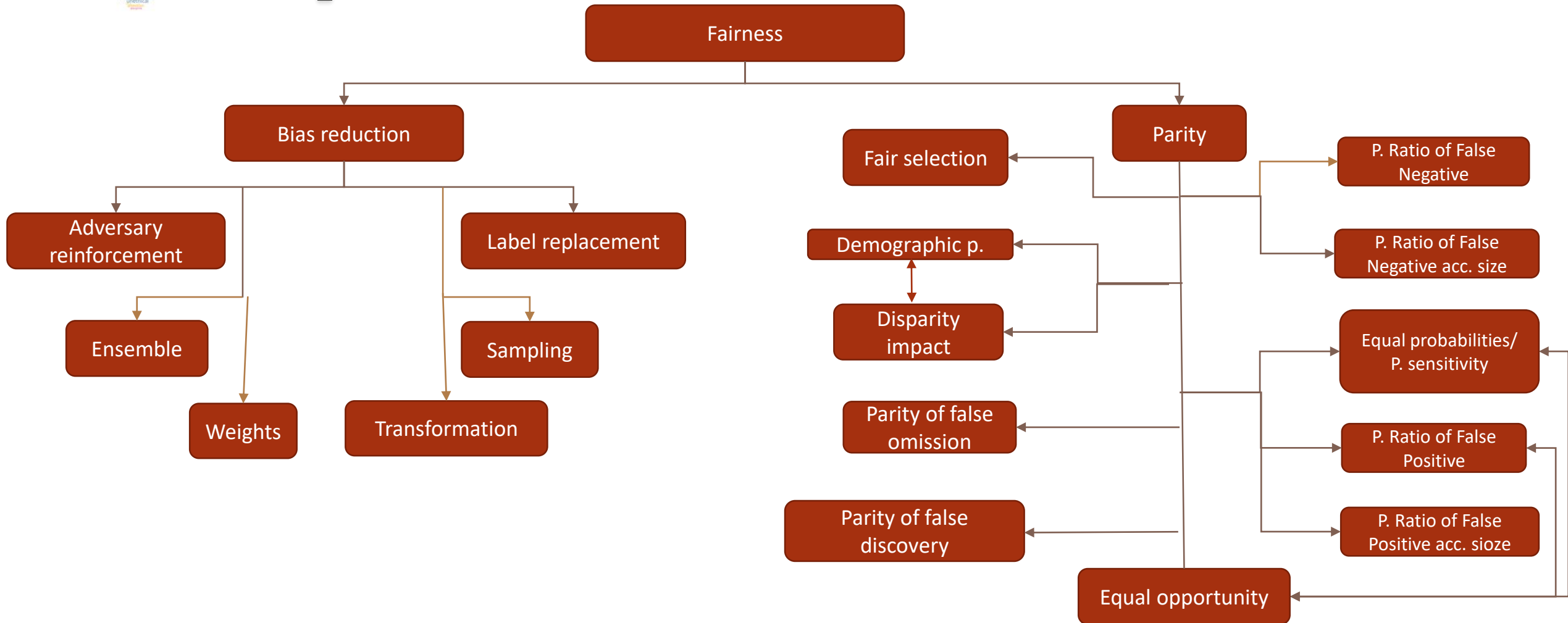


Fairness



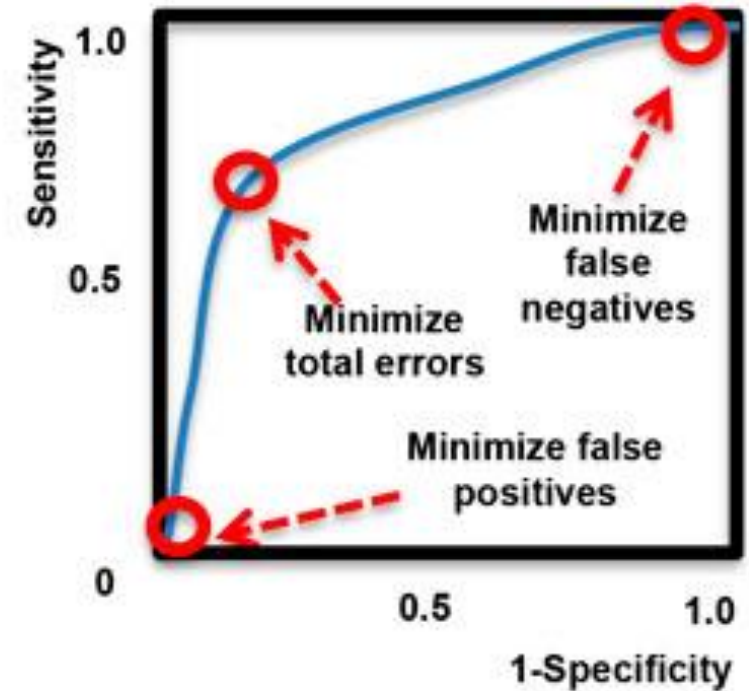
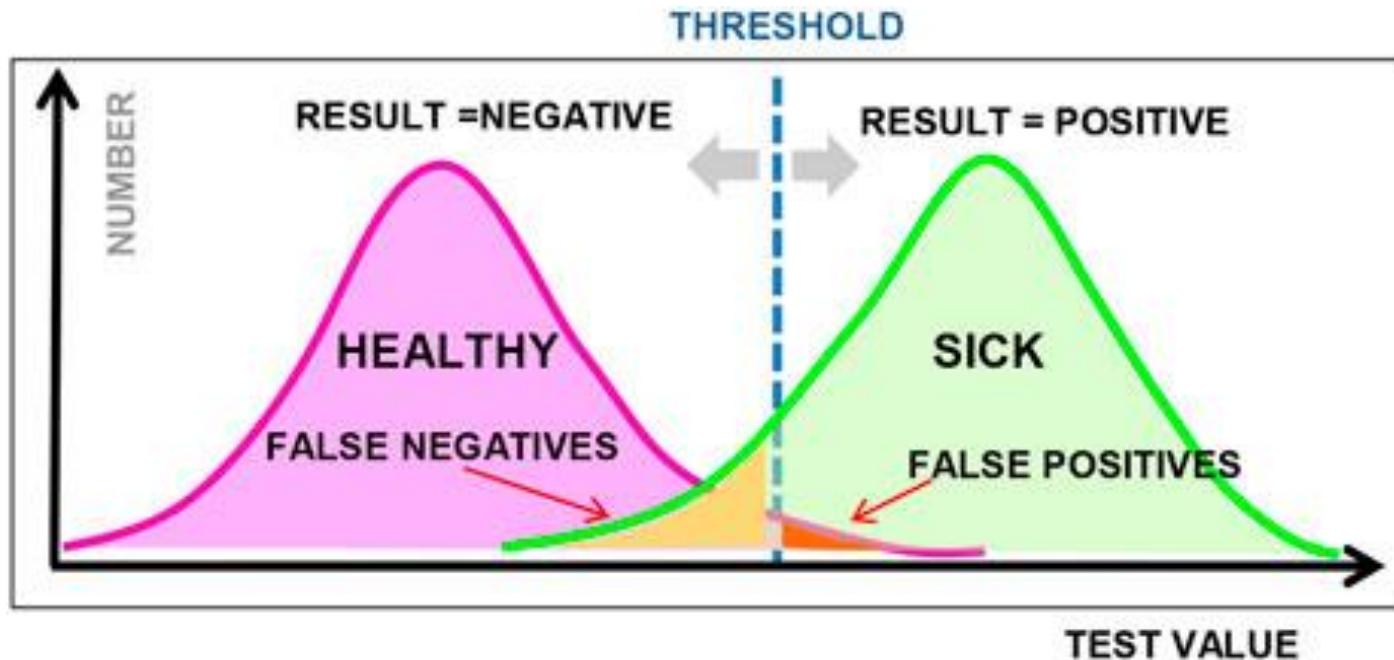


Fairness / Data fairness





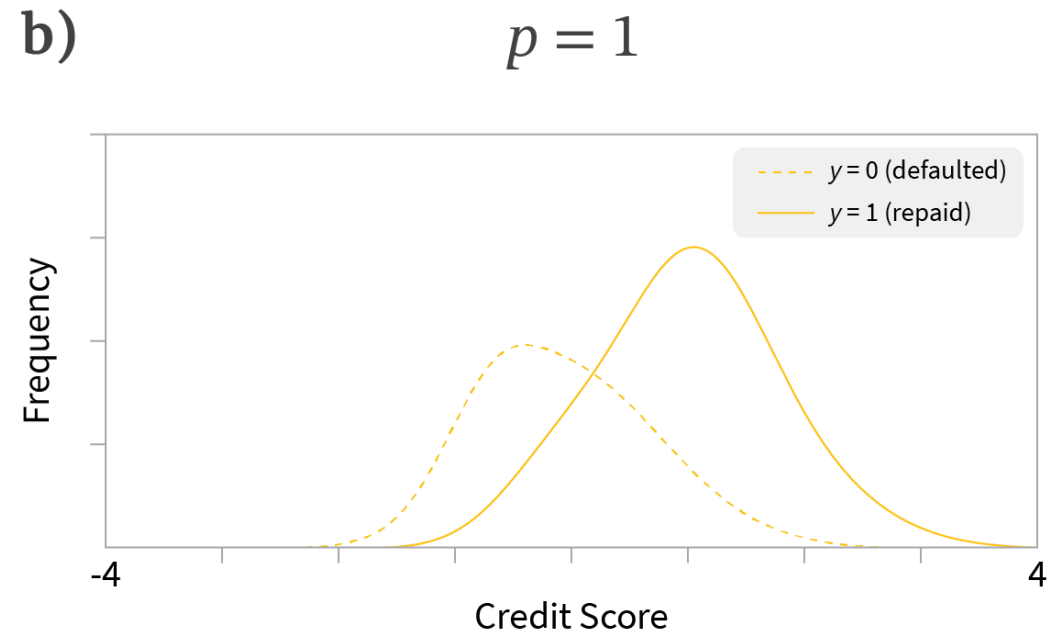
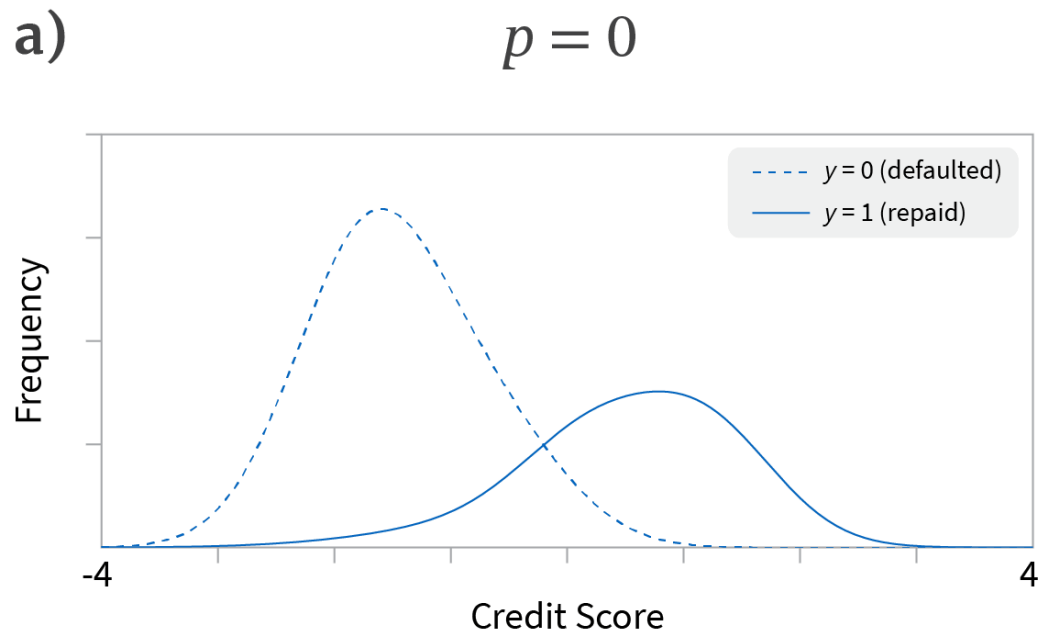
How we validate a ML model?



R.R. Fletcher. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front. Artif. Intell.* (2021)



Example of lack of impartiality: bank credit allocation

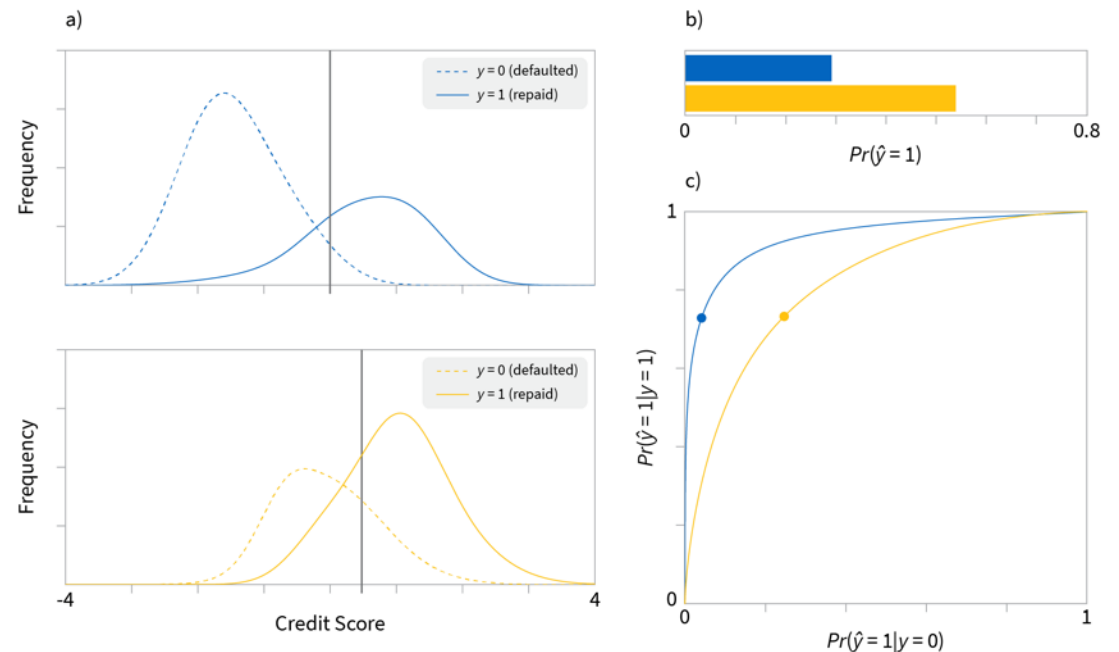
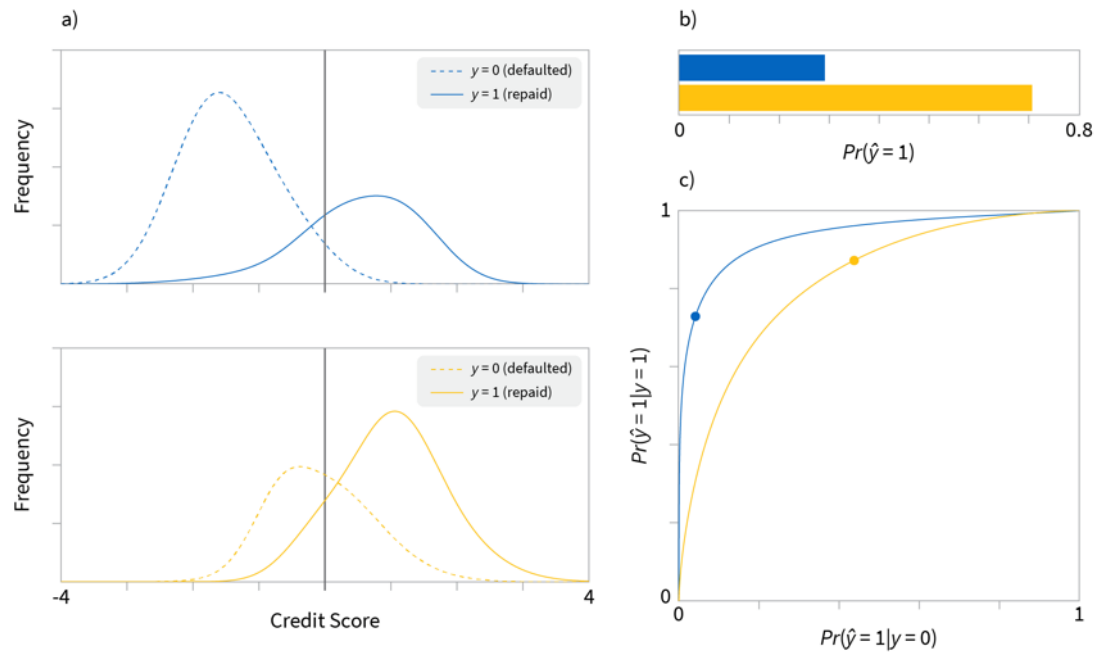




Example of lack of impartiality: bank credit allocation

Equalize population independent threshold (protected variable)

Equal opportunity, but disproportionate grouping for $y^{\wedge}=1$





Case study proving why a fair judicial algorithm is impossible





Auditability from a fairness point of view

Detect if protected variables are in use

Detect and evaluate lack of ethics in the model

Detect if there is an indirect influence with the protected variables.

Check what variables / instances are affecting the model

P. Adler *et al.*, «Auditing Black-box Models for Indirect Influence», *arXiv:1602.07043*, nov. 2016.

F. Preto, «Inclusive Machine Learning: addressing model fairness», *Towards Data Science*, ago. 12, 2020.

S. Lundberg, «Explaining Measures of Fairness», *Towards Data Science*, mar. 02, 2020.



How to proceed to avoid biases

The data used must represent "what should be" and not "what is".

- Control possible pre-acquired biases, ensuring that the data sample represents all cases equally and does not discriminate against any group.
- It is mandatory to follow a control in the collection and preparation of data, identifying and correcting a priori possible deficiencies

Impose and enforce some form of data governance.

- There is a social responsibility to regulate the model generation procedure.
- Requiring to achieve not only high predictive performance metrics (accuracy of the system in production), but complementing them with ethical measures, managing audits for the different algorithms.

Model evaluation should include an evaluation by social groups.

- Examine possible "protected" variables (sensitive characteristics such as gender, ethnicity or age) to ensure that quality metrics (accuracy) are unbiased across different population segments.

eXplanaible AI (XAI)

The goal of explainable AI

Today



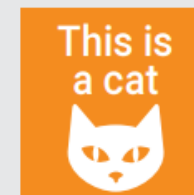
Training
Data



Learning
Process



Learned
Function



Output



User with
a Task

Tomorrow



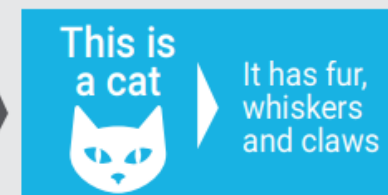
Training
Data



New Learning
Process



Explainable
Model



Explainable Interface

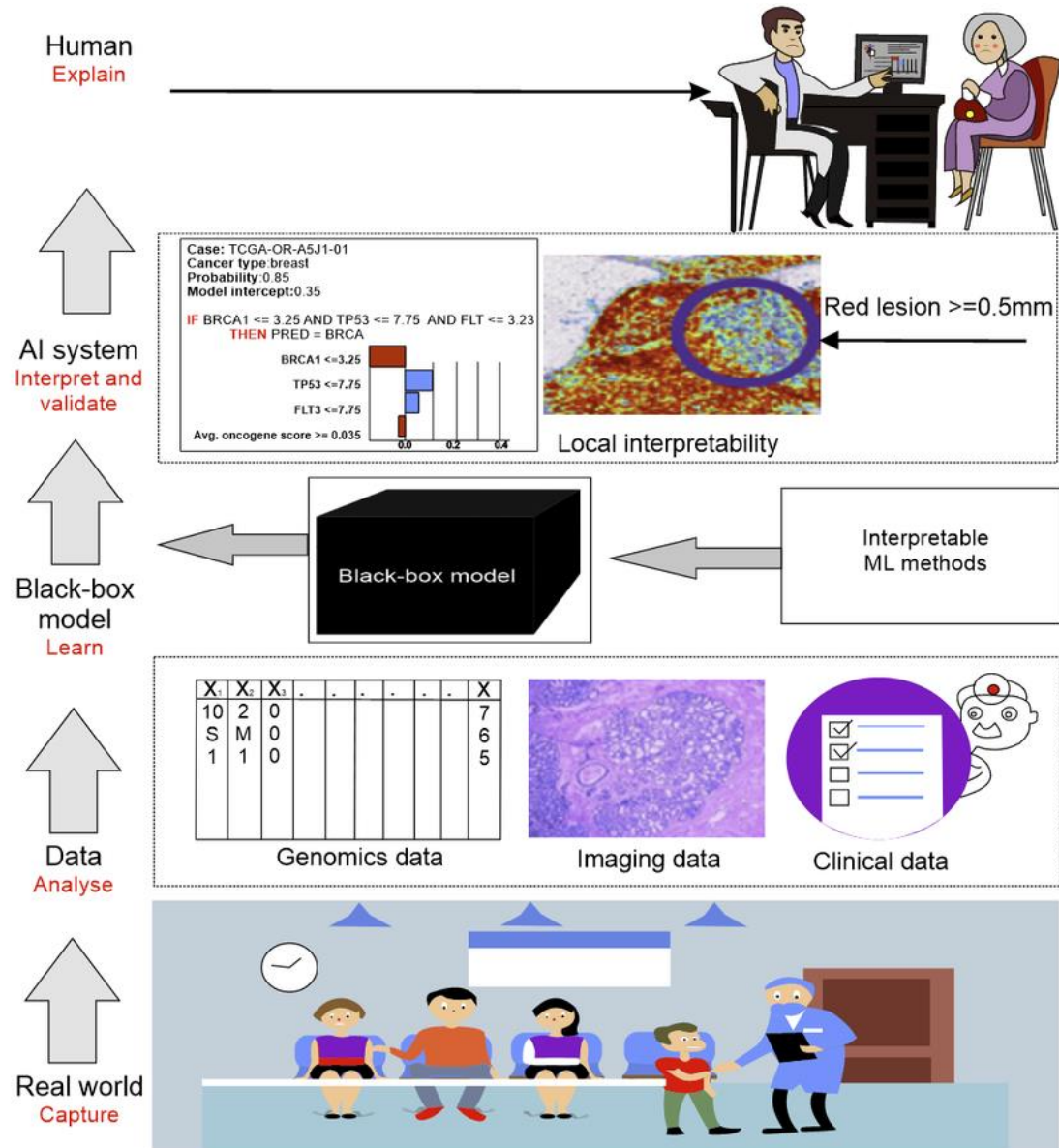


User with
a Task

XAI in breast cancer diagnosis

- Problems in bioinformatics are inherently multimodal.
- Balance between model accuracy / interpretation.
- Being able to answer:
 - "why do I have breast cancer?",
 - "how did the model arrive at this decision?",
 - "what biomarkers or factors are responsible?"

Md. Rezaul Karim, et al. Explainable AI for Bioinformatics: Methods, Tools, and Applications. Arxiv:2212.13261v1





Conceptos asociados a XAI

Transparency

- Simulability
- Decomposability
- Algorithmic transparency

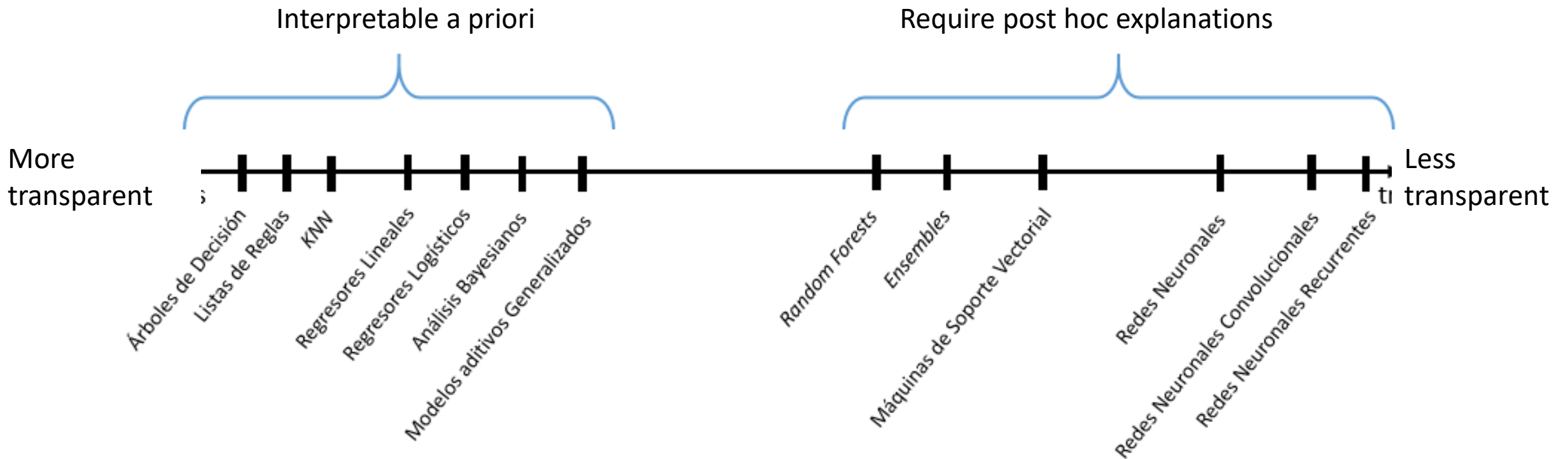
Post—hoc explainability

- Textual description
- Local models
- Visual models
- Analogies





XAI view of ML and AI models

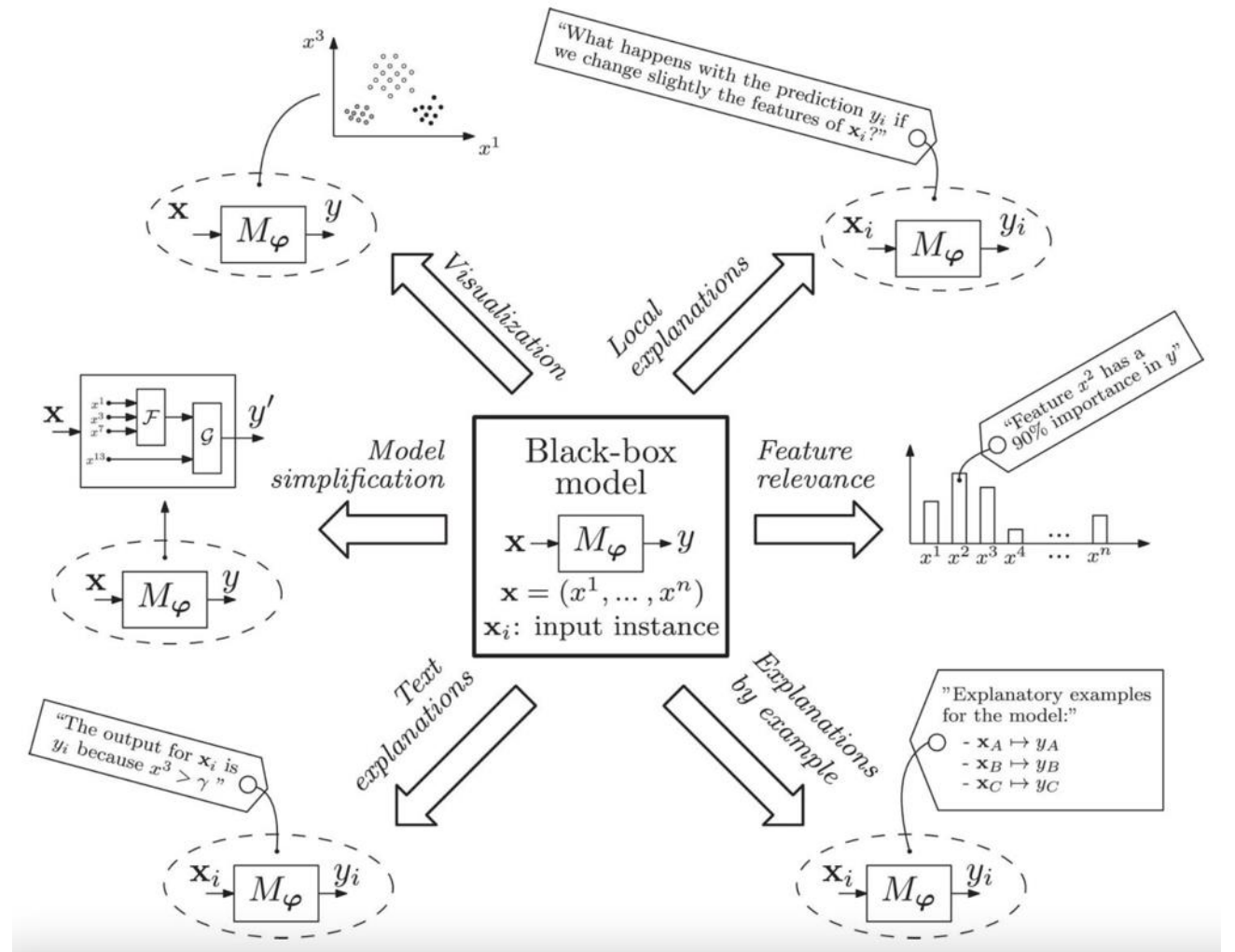


Different ways to improve the explanation

There are numerous approaches to extract the true knowledge from a predictor / ML model.

The preference for one option or another will depend on the context and the information to be extracted.

In all cases, it should be useful for the expert in making a final decision.





Interpretability strategies

Feature importance

- What changes to apply to change the instance category?
- What was the influence of the feature on the model output?

"Sensitivity" maps (heat maps)

- Illustrate variations in the importance of different features, using color to convey the weight of the pixel in a given prediction

Model visualization

- Patterns detected in an image
- Distribution of features in the dataset

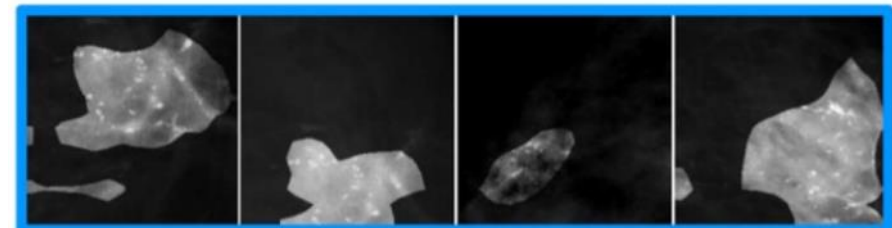
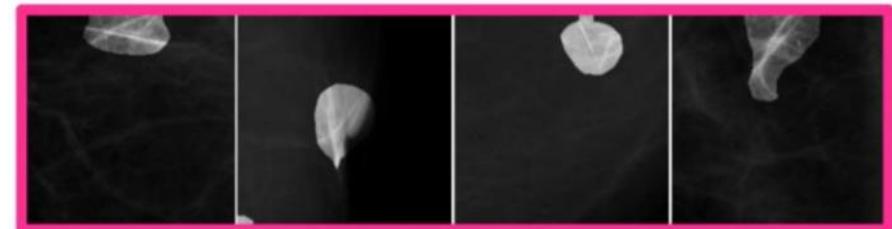
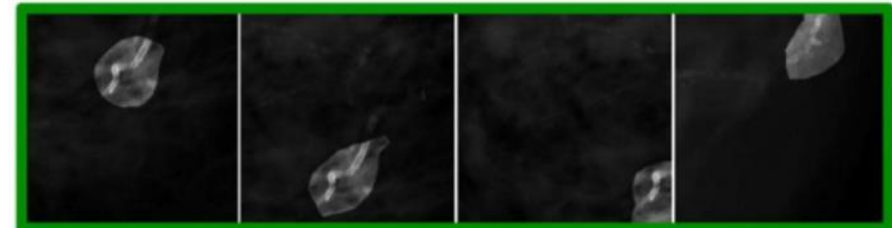
Surrogate models

- Generate simple model from original model predictions

Domain knowledge

Example-based explanation: factual / counterfactual

Top Activating Patches for each Target Unit



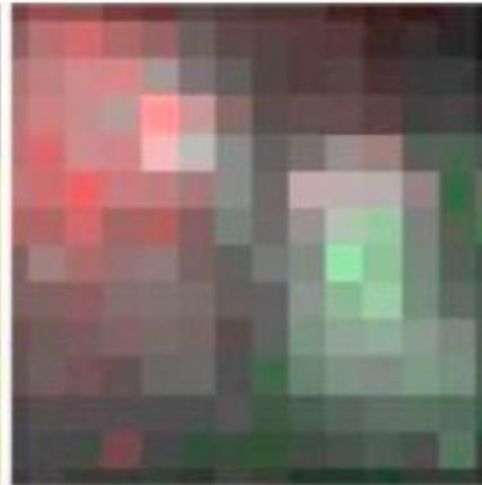


Example of a posteriori explainability



(a) Heatmap [168]

*What's the contribution of each single **pixel**?*



(b) Attribution [293]

*What **part** of an example is responsible for the network activating?*



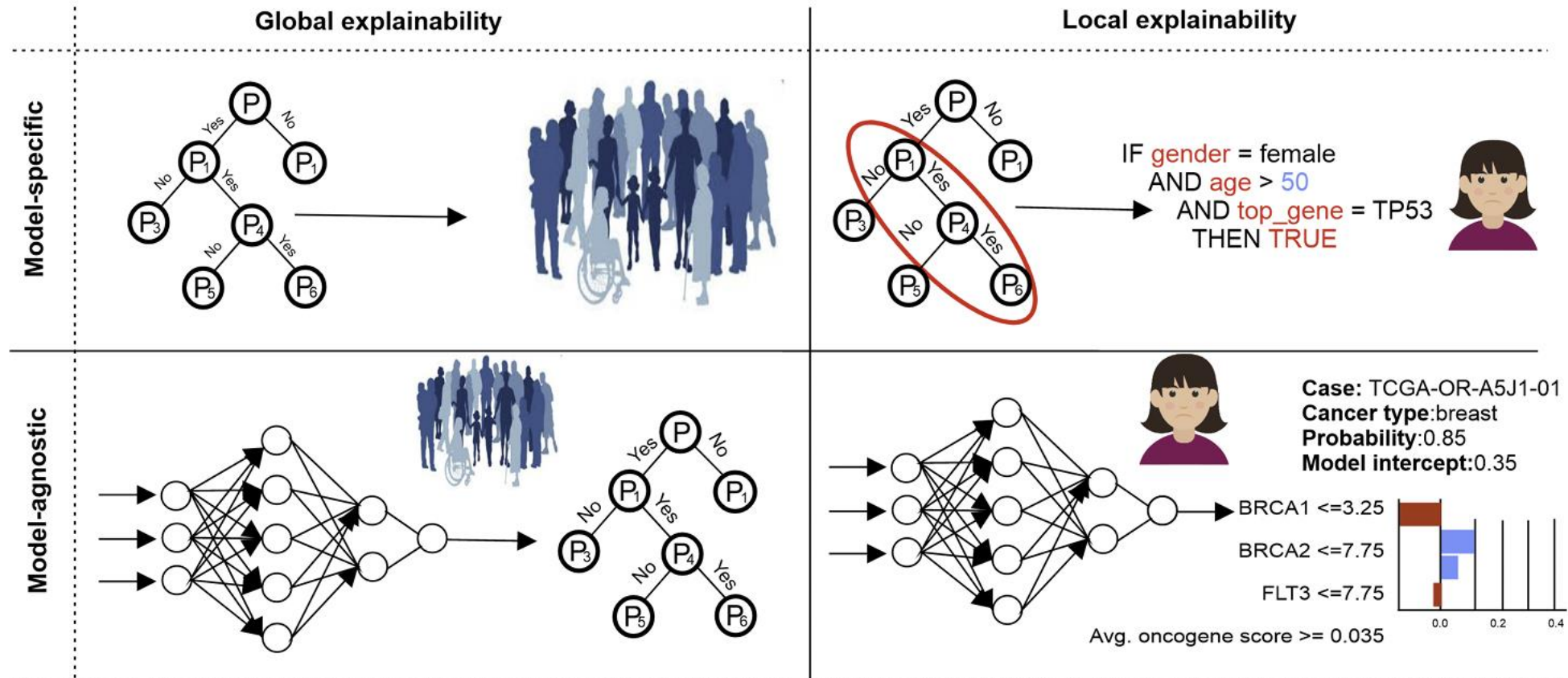
(c) Grad-CAM [292]

*What **regions** are important for predicting the concept?*

[\[Olah, Mordvintsev, Schubert, 2020\]](#)



Examples of explainability



Md. Rezaul Karim, et al. Explainable AI for Bioinformatics: Methods, Tools, and Applications. Arxiv:2212.13261v1





Auditability from “responsability”

- Establishing **quantity and quality** of training data: biases / parity (fairness).
- **Decentralized AI**: Integration of Machine Learning in the Blockchain.
- Control / anticipation of **attacks**: adversarial machine learning.
- Management of **concept drift** (data fracturing) within MLOps.
- Establishment of **roles** associated with the use:
 - Designer
 - Implementer
 - User

INDEX



1

Natural
Intelligence vs.
Artificial
Intelligence



2

What is ethics:
Ethics in
Artificial
Intelligence



3

Principles of
Responsible
Artificial
Intelligence



4

Artificial
Intelligence
Governance



5

Going deeper in
the study areas
on AI ethics



6

**Summary: AI
implications, AI
for Good,
suggestions**



Implications for Artificial Intelligence

Trustworthy AI / Ethics / Social good

- Humans must remain in control of AI
- How to design AI systems that are beneficial to society and robust?
- Interdisciplinary research: economics, law, security, biology, health, among others.

Challenges





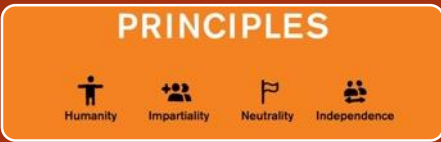
Artificial Intelligence for the good



Accessibility: equal opportunities for different users



Land: climate, water, agriculture and biodiversity.



Humanitarian: disaster recovery, needs of children, refugees and displaced persons, and human rights.



Cultural: simple access adapted to classical values.



Health: The initiative will focus on accelerating the search for possible treatments, increasing what we know about health and longevity, and reducing health inequality.



Short- and medium-term quality control issues in machine learning. (Clinical)

Disruptive change

- Has the system been tested in a variety of locations, underlying software architectures and populations?
- Training data match what we expect and do not contain biases?
 - Label quality: "truth" vs. Clinical opinion
 - Training set imbalance
 - It is applied to the same diagnosis context
- How is it going to be supervised and maintained to adjust the concept drift?

Insensibility to impact

- Does the system adapt its behavior when there are high impact results?
- Can the system identify outlier inputs and adjust its confidence consequently

Black box, security, automatization

- Are the systems predictions interpretable?
- Does it produce an estimation of the confidence
- Is prediction certainty communicated to physicians to avoid automation bias?

Reinforcement of outdated practices and self-fulfilling predictions

- How can be last minute changes adapted in the clinical practice?
- What aspects of the current clinical practices reinforces the system?

R. Challen, et al. Artificial intelligence, bias and clinical safety (2018)



Final recommendations

- Use a **human-centered** design approach: clarity control, diverse use cases, etc.
- Identify **multiple metrics** in evaluation: metrics for groups, fit for purpose (false alarms), accuracy vs. ethics metrics.
- Directly examine **raw data**: irregularities, fairness, etc.
- Understand **limitations of data set** and model: correlation vs. causation; scope of training and limitation; communicate these.
- **Test, test, test** (MLOps): unit test, integration, drift, requirements, etc.
- Monitor and **update** systems even after implementation.



Conclusions and lessons learned

AI Ethics is both a growing and developing branch.

There is a wide range of techniques appropriate for AI Ethics. Focus on data processing

There is still much debate about whether to use interpretable models or to explain black box models.

While the associated legislation is still developing, it is appropriate to adapt to good practice in AI Ethics.



THANK YOU VERY MUCH FOR YOUR ATTENTION

TO CONTACT ME FOR FURTHER QUESTIONS, PLEASE USE THE FOLLOWING FORM:



Main office

ETSIT, Dpt. CCIA, D16
C/Periodista Daniel Saucedo
18014. Granada



Alternative office

Centro empresas PTS
Av. Del Conocimiento 41,
18016. Granada



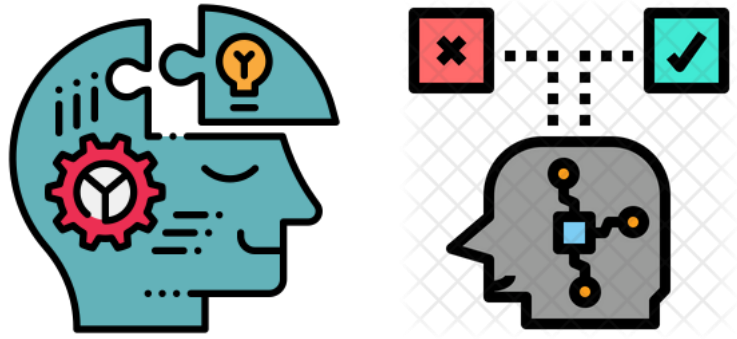
Phone

(+34) 958 240 079



Email / Web

alberto@decsai.ugr.es
<https://www.dasci.es>



Towards an Ethical Artificial Intelligence

Consolidating Transparency, Fairness and Accountability in Machine Learning Models

Alberto Fernández – Instituto DaSCI . Universidad de Granada

