ACM Summer School on User Modeling and Personalization in Urban Computing:

AI for social good

Dr Antonela Tommasel

ISISTAN, CONICET-UNICEN, Argentina

CONICET

ISISTAN

Table of contents

1. **AI and social good?**

2. **AI areas of action**

3. **Responsible AI practices**

Table of contents

1. **AI and social good?**

2. AI areas of action

3. Responsible AI practices

AI and social good?

AI is not a silver bullet, but it could help tackle some of the world’s most challenging social problems.

What can be done?

- The **United Nations’ Sustainable Development Goals** are among the best-known and most frequently cited societal challenges.

Life below water
Affordable and clean energy
Clean water and sanitation
Responsible consumption and production
Sustainable cities and communities
Gender equality
Revitalize the global partnership for sustainable development.
Zero hunger

Decent work and economic growth
Climate action
Reduced inequalities
Industry, innovation, and infraestructura
No poverty
Life on land
Quality education
Peace, justice, and strong institutions
Good health and well-being

AI areas of action

- The UN goals set different areas in which AI could help!

- Public and social sector
- Security and justice
- Crisis response
- Economic empowerment
- Education
- Infraestructure
- Information verification and validation
- Health and hunger
- Equality and inclusion
- Environment

McKinsey Global Institute Analysis

AI areas of action

- The UN goals set different areas in which AI could help!

- Initiatives related to efficiency and the effective management of public- and social-sector entities, including:
  - Strong institutions.
  - Transparency.
  - Financial management.

- For example, AI can be used to identify **tax fraud** using alternative data such as browsing data, retail data, or payments history.

Public and social sector

Revitalize the global partnership for sustainable development.

Good health and well-being

AI areas of action

Digital Citizenship

Public and social sector

- Democracy and the formal political process dependent on the effective communication about public issues amongst citizens.

- E-participation involves the extension and transformation of participation in societal democratic processes mediated by technologies.

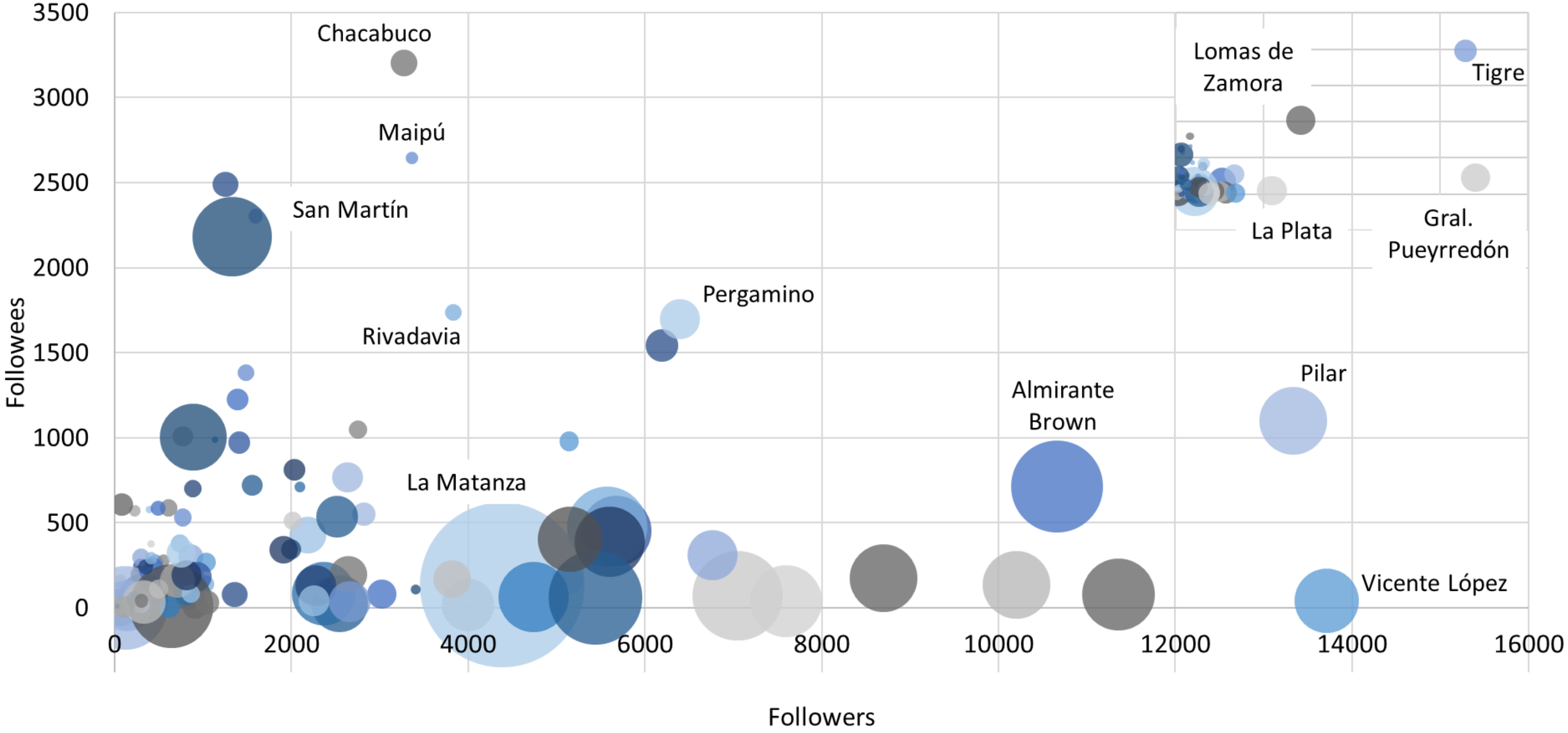
- Support active citizenship with the latest technology developments, increasing access to and availability of participation to promote fair and efficient societies and governments.

- What can we do with this?
  - Study the relationship of the government with the citizens.
  - Study the characteristics of citizens.
  - Create citizens' profiles.
  - Estimate the engagement cycle of citizens.
  - Try to foster the engagement of citizens.

AI areas of action

Digital Citizenship

Public and social sector

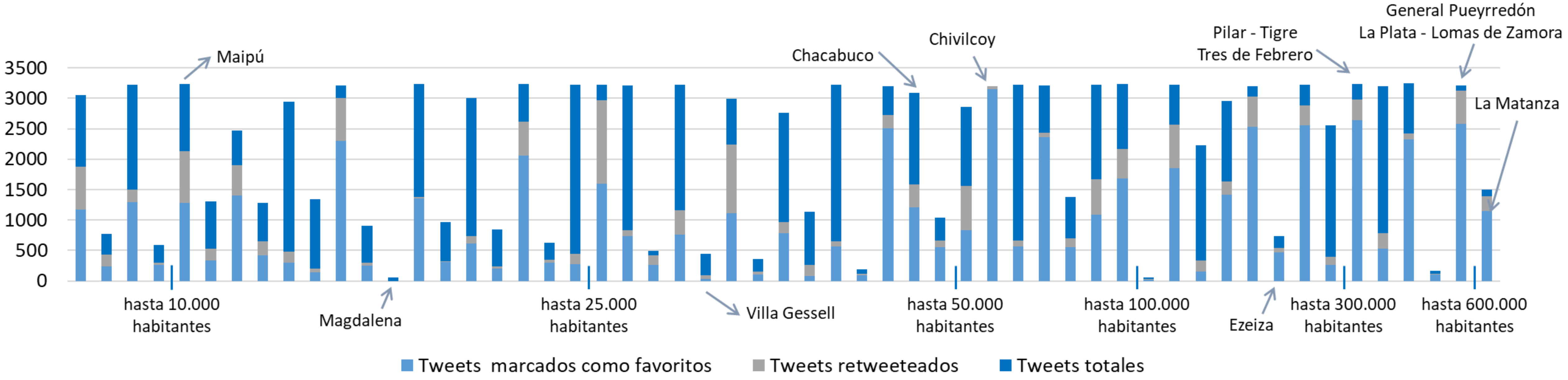


Number of followees and followers in Twitter according to number of habitants

AI areas of action

Digital Citizenship

Public and social sector



AI areas of action

- The UN goals set different areas in which AI could help!

- These are specific crisis-related challenges, such as responses to natural and human made disasters in search and rescue missions, as well as the outbreak of diseases.

- For example:
  - Using AI on satellite data to map and predict the progression of wildfires and thereby optimize the response of firefighters.

  - Drones with AI capabilities can also be used to find missing persons in wilderness areas.

Crisis response

Peace, Justice and strong institutions
Good health and well-being

AI areas of action

- The UN goals set different areas in which AI could help!

- Emphasis on currently vulnerable populations.
- Opening access to economic resources and opportunities:
  - Jobs.
  - Market information.
- For example,
  - AI can be used to detect plant damage early through low-altitude sensors (including smartphones and drones), to improve yields for small farms.
  - AI can be used to determine the body condition score from cows.

Economic empowerment

Zero hunger
Decent work and economic growth
No poverty
Peace, justice and strong institutions

AI areas of action

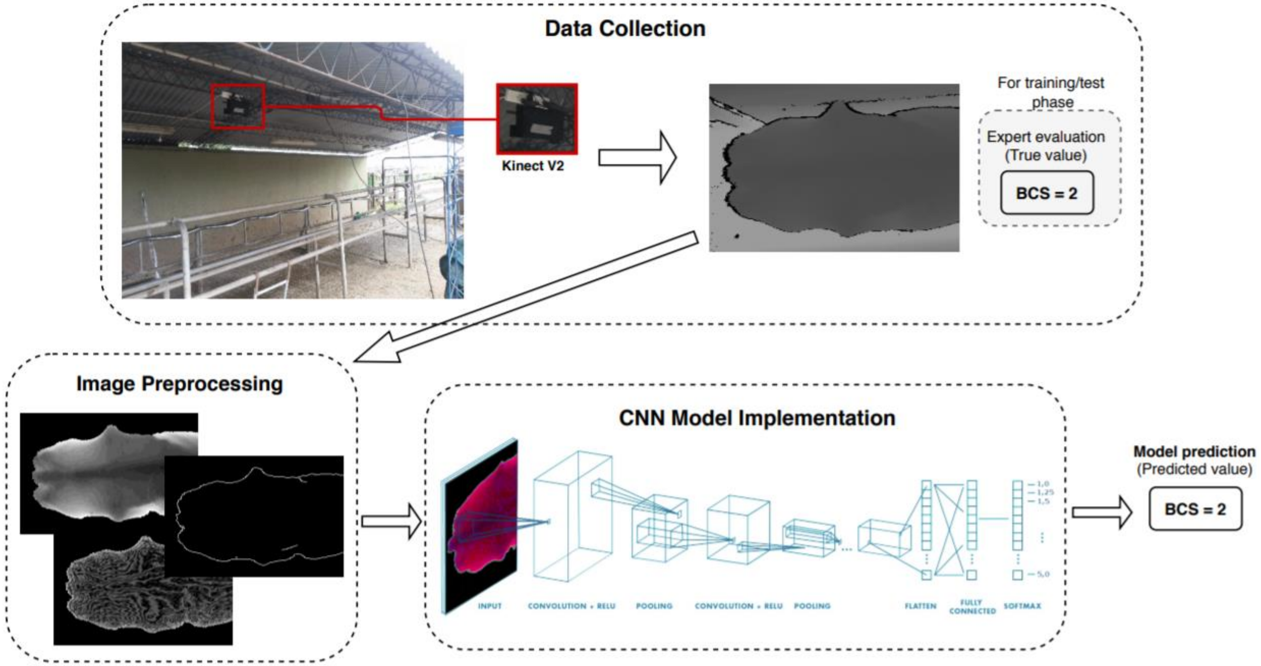
Estimating body condition of cows

Economic empowerment

- Body Condition Score is especially important for dairy cows as it is not only a measurement of **obesity** degree, but also a suitable assessment of **feeding management** according to each stage of lactation, which heavily influences milk production, reproduction, and cow health.

- Usually a time-consuming manual task performed by experts.

- Applied a novel CNN-based model to estimate BCS on cows from depth images.



Rodríguez Alvarez, J., Arroqui, M., Mangudo, P., Toloza, J., Jatip, D., Rodriguez, J.M., Teyseyre, A., Sanz, C., Zunino, A., Machado, C. and Mateos, C., 2019. Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. *Agronomy*, *9*(2), p.90.

AI areas of action

- The UN goals set different areas in which AI could help!

  - Maximizing student achievement and improving teachers’ productivity.

  - For example,
    - Adaptive-learning technology could be used for recommending content to students on past success and engagement with the material.

    - Student profiles could be created to adapt the course methodology to personality or learning style.

Education

Quality education

AI areas of action

- The UN goals set different areas in which AI could help!

- Sustaining biodiversity.

- Combating the depletion of natural resources, pollution.

- Climate change.

- For example,
  - The Rainforest Connection, a Bay Area non-profit, uses Tensor Flow in conservancy efforts to detect illegal logging in vulnerable forest areas by analysing audio-sensor data.

  - Global Fishing Watch. Tracks and monitors the global comercial fishing fleet.

Environmental

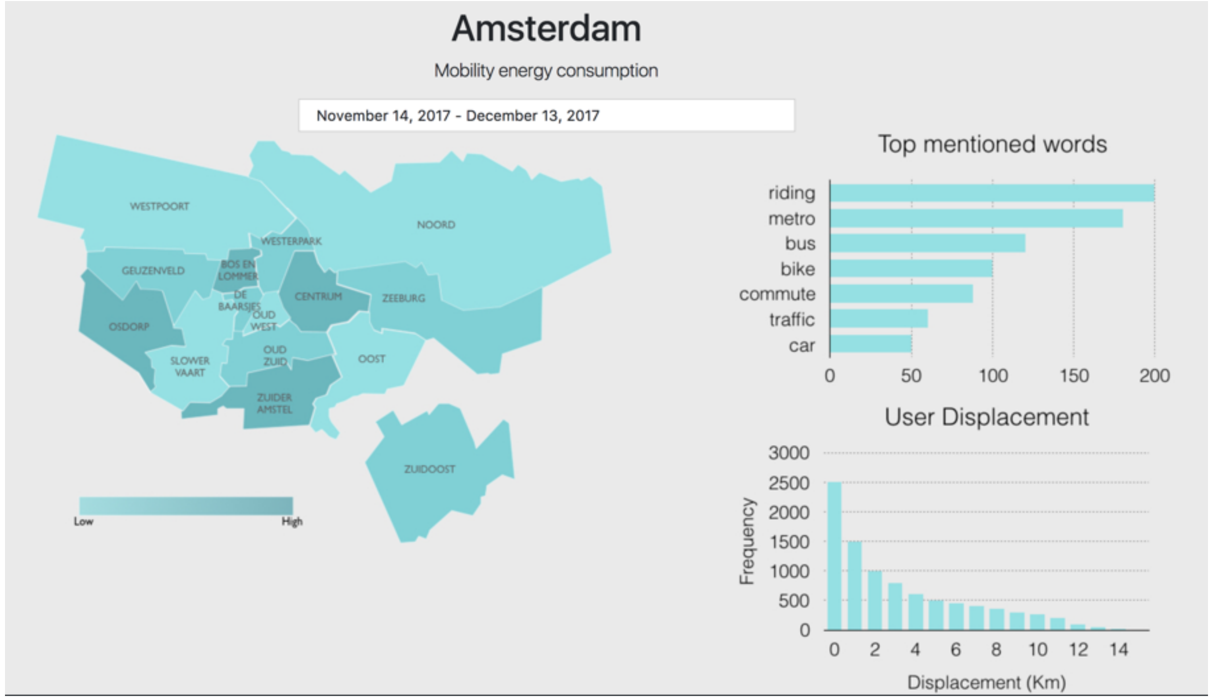
- Life below water
- Affordable and clean energy
- Climate action
- Life on land

AI areas of action

Energy consumption from social media

Environmental

- Traditional sources of information about energy consumption, such as smart devices and surveys, can be costly to deploy, may lack contextual information or have infrequent updates.
- Extracting energy consumption-related information from user-generated content.
- A pipeline that helps identify energy-related content in Twitter posts and classify it into four categories:
  - Dwelling.
  - Food.
  - Leisure.
  - Mobility.
- It aggregates the tweets into various spatial units (e.g., neighbourhoods, census tracts) and into hourly time slots.



Mauri, A., Psyllidis, A. and Bozzon, A., 2018, April. Social Smart Meter: Identifying energy consumption behavior in user-generated content. In *Companion Proceedings of the The Web Conference 2018* (pp. 195-198). International World Wide Web Conferences Steering Committee.

AI areas of action

- The UN goals set different areas in which AI could help!
- Equality.
- Inclusion.
- Self-determination.
- Reducing bias based on race, religion, citizenship…
- For example,
  - MIT Media Lab and Autism Glass involves using AI to automatically recognize emotions and to provide social cues to help individuals along the autism spectrum interact in social environments.

Equality and inclusion

Gender Equality
Decent work and economic growth
Reduced inequalities

AI areas of action

- The UN goals set different areas in which AI could help!

- Early-stage diagnosis.

- Optimized food distribution.

- For example,
  - At the University of Heidelberg and Stanford University have created a disease-detection AI system using images of skin lesions to determine if they are cancerous.

  - AI-enabled wearable devices can already detect people with potential early signs of diabetes using heart-rate sensor data.

  - Sedentary detection.

Health and hunger

Zero hunger
Good health and well-being

AI areas of action

- The UN goals set different areas in which AI could help!

- Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** online to large audiences willing to believe the fictions and spread the word

- Facilitating the provision, validation, and recommendation of helpful, valuable, and reliable information to all.

- Focuses on filtering or counteracting misleading and distorted content, including false and polarizing information disseminated through Internet and social media.

- A threat to the access to reliable and **trustworthy** information and the establishment of **reliable** social relations.

Information validation and verification

Peace, justice and strong institutions

Additional “real life” consequences!

- Mob killings
- Stock changes
- Health
- Political manipulation

AI areas of action

Fake news, social bots & more

Information validation and verification

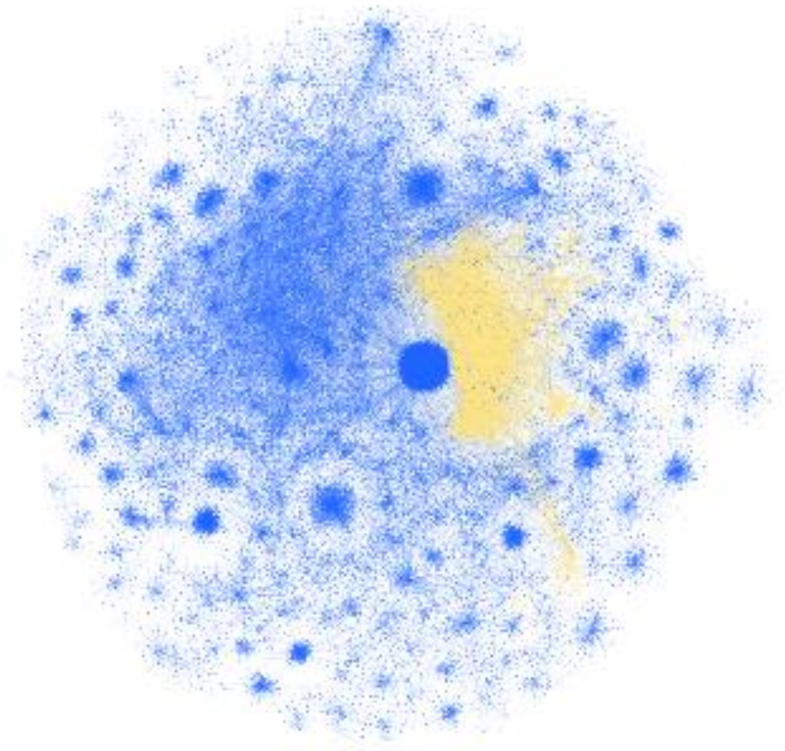
- Social media are vulnerable to deceptive social bots, which can impersonate humans to amplify misinformation and manipulate opinions.
- Little is known about the large-scale consequences of such pollution operations.
- Some social bots are designed with benign intentions and serve useful purposes, but they also have many harmful applications:
  - Amplify misinformation.
  - Create the appearance that people or opinions are more popular than they are.
  - Influence public opinion.
  - Commit financial fraud.
  - Infiltrate vulnerable communities.
  - Disrupt communication.
- Some social bots behave in ways that cannot be distinguished by those of humans on an individual basis because automation is mixed with or copied from human behaviour.

Lou, X., Flammini, A. and Menczer, F., 2019. Information Pollution by Social Bots. *arXiv preprint arXiv:1907.06130*.

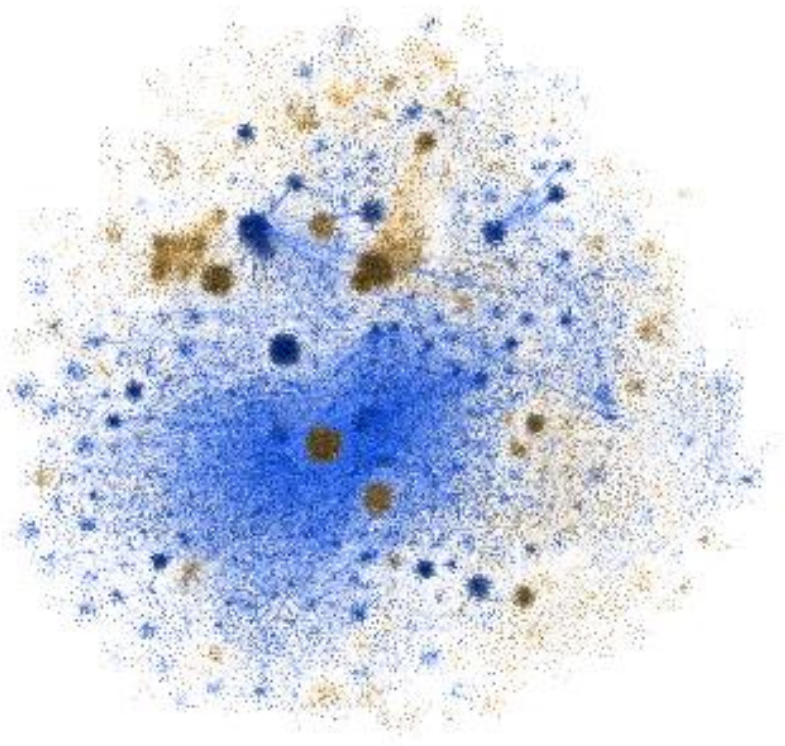
AI areas of action

Fake news, social bots & more

Information validation and verification



chemtrails conspiracies (yellow) mix with conversations about the sky

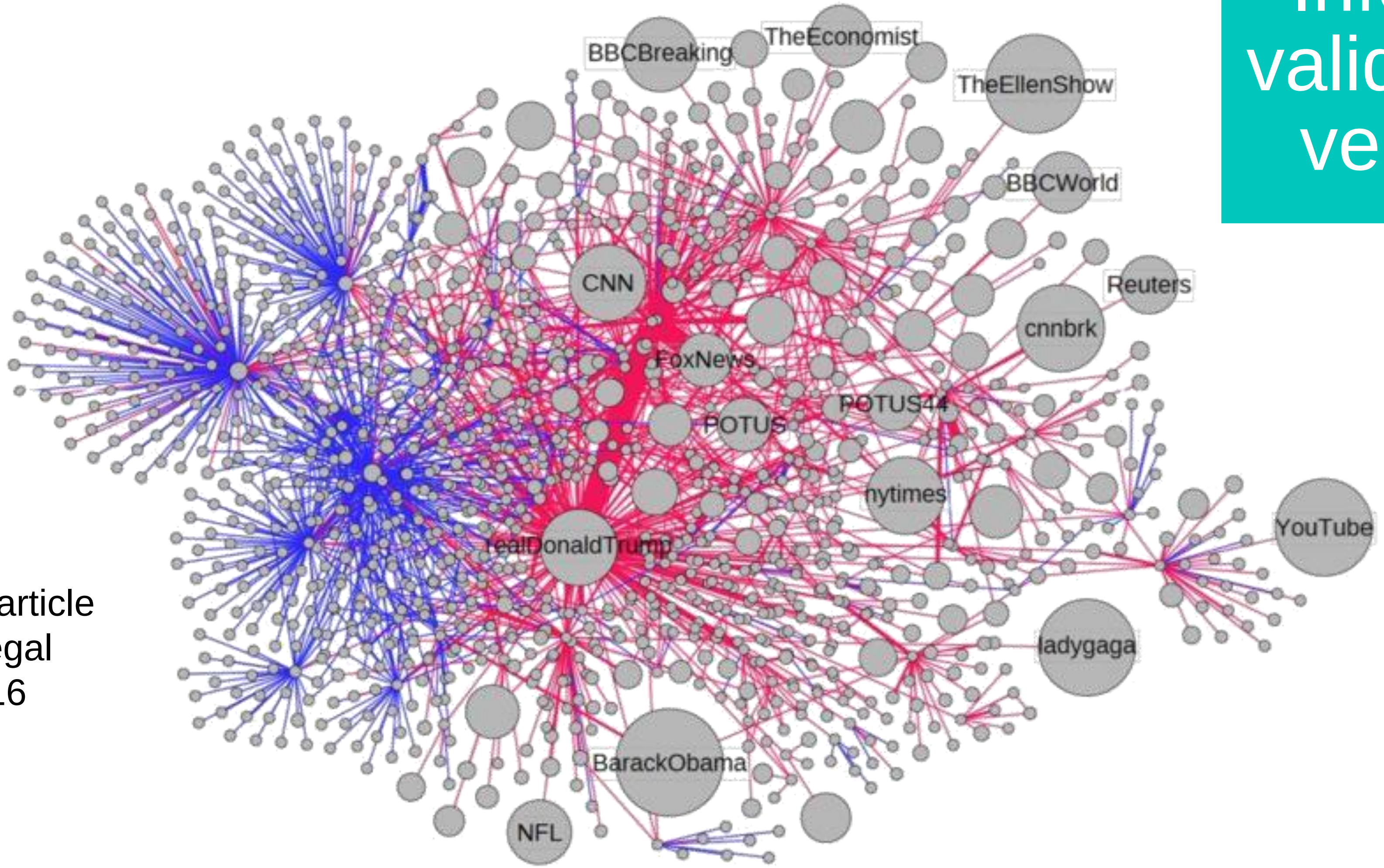


antivax campaigns (brown) penetrate discussions about the flu

AI areas of action

Fake news, social bots & more

Information validation and verification



Visualization of the spread through social media of an article falsely claiming 3 million illegal immigrants voted in the 2016 presidential election.

AI areas of action

- The UN goals set different areas in which AI could help!

- Promotes the public good in:
  - Energy.
  - Water and waste management.
  - Transportation.
  - Real estate.
  - Urban planning.

- For example:
  - Traffic-light networks can be optimized using real-time traffic
  - camera data and sensors to maximize vehicle throughput.
  - AI can also be used to schedule predictive maintenance of public transportation systems.
  - Crowd-sensing could be used for determining when to pick up waste.

Infrastructure management

Affordable and clean energy

Clean water and sanitation

Sustainable cities and communities

Reduced inequalities

Industry, innovation and infrastructure

No poverty

AI areas of action

Attendees in an event

Infrastructure management

- City-scale events attract large amounts of attendees in temporarily re-purposed urban environments.

- The real-time measurement of the density of attendees stationing in the event, such as crowd management, emergency support, and quality of service evaluation.

- Sensing or communication infrastructures can be deployed to estimate the number of attendees currently occupying an area.
  - These technologies is hindered by their cost or sensing resolution.
  - Social media data can provide a real-time and semantically rich insight into attendees' behaviour during city-scale events.

- How micro-posts harvested from social media can be used during city-scale events to estimate the density of attendees in a given terrain?

- Three classes of density estimation strategies:
  - Geo-based.
  - Speed-based.
  - Flow-based.

Gong, V.X., Yang, J., Daamen, W., Bozzon, A., Hoogendoorn, S. and Houben, G.J., 2018. Using social media for attendees density estimation in city-scale events. *IEEE Access*, *6*, pp.36325-36340.

AI areas of action

- The UN goals set different areas in which AI could help!

- Focuses on security, policing, and criminal-justice issues.
  - Preventing crime and other physical dangers.
  - Tracking criminals.
- Mitigating bias in police forces.

- For example,
  - Using AI and IoT devices to create solutions that to help firefighters determine safe paths through burning buildings.

  - Design and simulate evacuation plans.

Security and justice

Sustainable cities and communities
Reduced inequalities
Peace, justice, and strong institutions

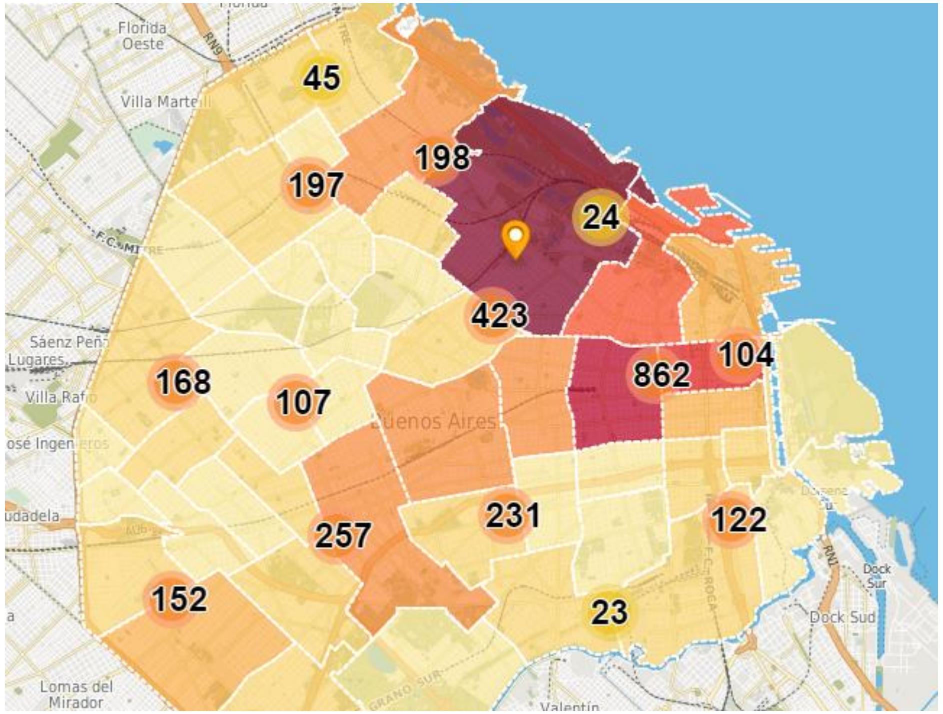
AI areas of action

Crime map

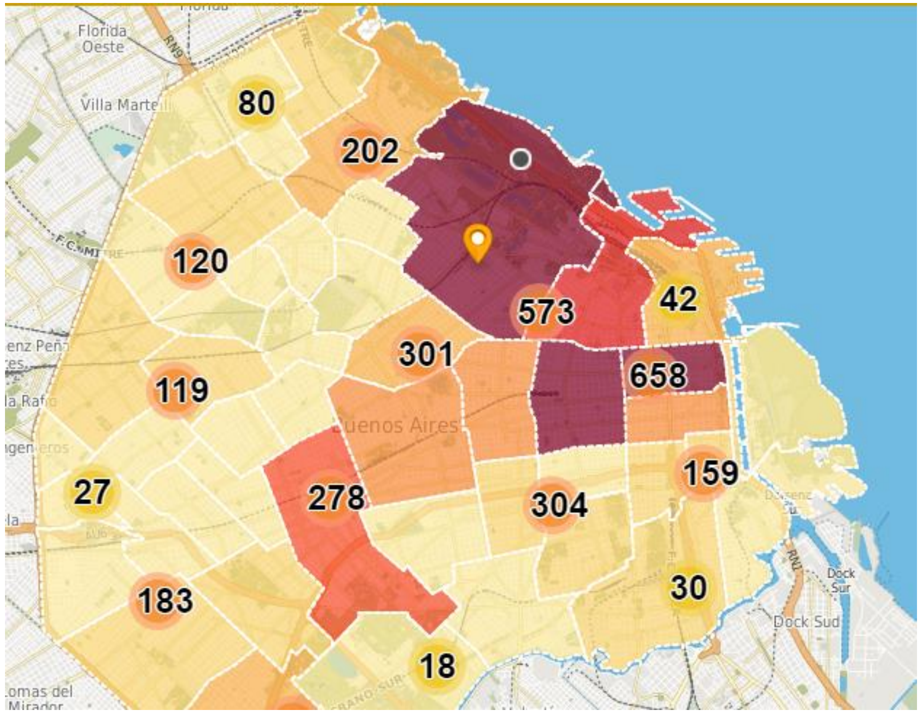
Buenos Aires Ciudad | Mapa del Delito

Security and justice

September 2018



December 2018



Limitations for AI use

Critical barriers for most domains

- Data availability
- Data quality
- High-level AI-expertise availability
- Regulatory limitations
- Data volume
- Data labelling
- Access to computing capacity

Contextual challenges

- Data integration
- Access to technology
- Privacy concerns
- Receptiveness

Limitations for AI use

Critical barriers for most domains

- **Data availability**
- Data quality
- High-level AI-expertise availability
- Regulatory limitations
- Data volume
- Data labelling
- Access to computing capacity

Contextual challenges

- **Data integration**
- Access to technology
- Privacy concerts
- Receptiveness

- Much of the data essential or useful for social-good applications are in **private hands** or in public institutions that might not be willing to share their data.

- These data owners include:
  - Telecommunications and satellite companies.
  - Social-media platforms;
  - Financial institutions (e.g. credit histories).
  - hospitals, doctors, and other health providers.
  - Governments (e.g. tax information for private individuals).

- Why is data difficult to obtain?
  - Regulations on data use.
  - Privacy concerns.
  - Bureaucratic inertia.
  - Data has business value → **commercial use!**

Limitations for AI use

Critical barriers for most domains

- Data availability
- Data quality
- **High-level AI-expertise availability**
- Regulatory limitations
- Data volume
- Data labelling
- Access to computing capacity

Contextual challenges

- Data integration
- Access to technology
- Privacy concerts
- **Receptiveness**

- Not only expertise in building the models but also on interpreting the results.
- Even if a model achieves a desired level of accuracy on test data, new or unanticipated failure cases often appear in real-life scenarios.
  - Need someone to “translate” it.
  - Without “translation” the model could be trusted too much!

Usually, models do not work on the 100% of the cases!

Table of contents

1. AI and social good?

2. AI areas of action

3. **Responsible AI practices**

Risks in AI & responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.

Risks in AI & responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.

AI tools and techniques can be misused so principles for their use must be established.

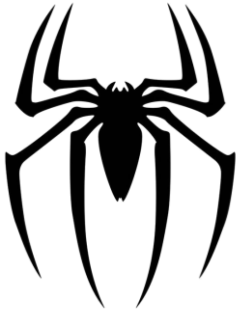
Risks in AI & responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.

AI tools and techniques can be misused so principles for their use must be established.

It is also raising new questions about the best way to build the solutions.

“with great power comes great responsibility”



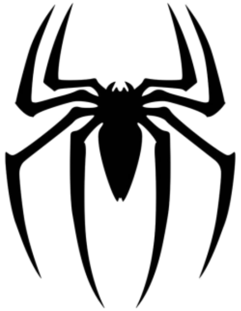
Risks in AI & responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.

AI tools and techniques can be misused so principles for their use must be established.

It is also raising new questions about the best way to build the solutions.

“with great power comes great responsibility”



How to be responsible?

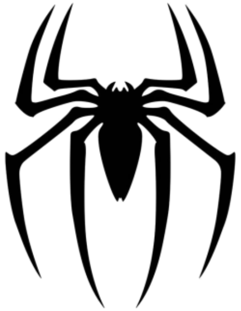
Risks in AI & responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.

AI tools and techniques can be misused so principles for their use must be established.

It is also raising new questions about the best way to build the solutions.

“with great power comes great responsibility”



How to be responsible?

Fairness

Interpretability

Privacy

Security

Responsible AI practices

Fairness

- Beyond recommending books and television shows, AI systems can be used for more **critical** tasks.
- Systems have the potential to be **fairer** and **more inclusive** at a broader scale than decision-making processes based on **ad hoc rules or human judgments.**
- AI solutions can unintentionally harm the very people they are supposed to help.
- The **risk** is that any **unfairness** in such systems can also **have a wide-scale impact**.
- Models can even be used to identify some of the conscious and unconscious human biases and barriers to inclusion that have developed and perpetuated throughout history, bringing about positive change.

Responsible AI practices

Fairness

- Beyond recommending books and television shows, AI systems can be used for more **critical** tasks.
- Systems have the potential to be **fairer** and **more inclusive** at a broader scale than decision-making processes based on **ad hoc rules or human judgments.**
- AI solutions can unintentionally harm the very people they are supposed to help.
- The **risk** is that any **unfairness** in such systems can also **have a wide-scale impact**.
- Models can even be used to identify some of the conscious and unconscious human biases and barriers to inclusion that have developed and perpetuated throughout history, bringing about positive change.

- Presence and severity of a medical condition.
- Matching people to jobs and partners.
- Identifying if a person is crossing the street.

AI may perpetuate and aggravate existing prejudices and social inequalities, affecting already-vulnerable populations and amplifying existing cultural prejudices.

It is critical to work towards systems that are fair and inclusive for all.

problematic historical data, including unrepresentative or inaccurate sample sizes

Responsible AI practices

Fairness

- But… not so easy

1. Models learn from existing data collected from the real world, and so **an accurate model may learn or even amplify problematic pre-existing biases** in the data based on race, gender, religion or other characteristics.

2. Does not matter how much training and testing we do, it is a challenge to ensure that a system will be fair across every possible situation.

3. There is no standard definition of fairness, whether decisions are made by humans or machines.

    - Identifying appropriate fairness criteria for a system requires accounting for user experience, cultural, social, historical, political, legal, and ethical considerations.

Responsible AI practices

Fairness

- **But… not so easy**

A job-matching system might learn to favour male candidates for CEO interviews, or assume female pronouns when translating words like “nurse” or “babysitter” into gendered languages (such as Spanish), because that matches historical data.

1. Models learn from existing data collected from the real world, and so **an accurate model may learn or even amplify problematic pre-existing biases** in the data based on race, gender, religion or other characteristics.

2. Does not matter how much training and testing we do, it is a challenge to ensure that a system will be fair across every possible situation.

3. There is no standard definition of fairness, whether decisions are made by humans or machines.

    - Identifying appropriate fairness criteria for a system requires accounting for user experience, cultural, social, historical, political, legal, and ethical considerations.

Responsible AI practices

Fairness

- **But… not so easy**

1. Models learn from existing data collected from the real world, and so **an accurate model may learn or even amplify problematic pre-existing biases** in the data based on race, gender, religion or other characteristics.

2. Does not matter how much training and testing we do, it is a challenge to ensure that a system will be fair across every possible situation.

3. There is no standard definition of fairness, whether decisions are made by humans or machines.

   - Identifying appropriate fairness criteria for a system requires accounting for user experience, cultural, social, historical, political, legal, and ethical considerations.

A job-matching system might learn to favour male candidates for CEO interviews, or assume female pronouns when translating words like “nurse” or “babysitter” into gendered languages (such as Spanish), because that matches historical data.

For example, a speech recognition system that was trained on US adults may be fair and inclusive in that context. When used by teenagers, however, the system may fail to recognize evolving slang words or phrases.

Use of the system after launch can reveal unintentional, unfair blind spots that were difficult to predict.

Responsible AI practices

Fairness

- **But… not so easy**

1. Models learn from existing data collected from the real world, and so **an accurate model may learn or even amplify problematic pre-existing biases** in the data based on race, gender, religion or other characteristics.

2. Does not matter how much training and testing we do, it is a challenge to ensure that a system will be fair across every possible situation.

3. There is no standard definition of fairness, whether decisions are made by humans or machines.

   - Identifying appropriate fairness criteria for a system requires accounting for user experience, cultural, social, historical, political, legal, and ethical considerations.

A job-matching system might learn to favour male candidates for CEO interviews, or assume female pronouns when translating words like “nurse” or “babysitter” into gendered languages (such as Spanish), because that matches historical data.

For example, a speech recognition system that was trained on US adults may be fair and inclusive in that context. When used by teenagers, however, the system may fail to recognize evolving slang words or phrases.

Use of the system after launch can reveal unintentional, unfair blind spots that were difficult to predict.

Even for situations that seem simple, people may disagree about what is fair, and it may be unclear what point of view should dictate policy, especially in a global setting.

Responsible AI practices

Fairness: Recommended practices

- It is important to identify whether or not machine learning can help provide an adequate solution to the specific problem at hand.

- If it can, just as there is **no single "correct" model** for all tasks, there is **no** single technique that ensures fairness **in every situation.**

Design models using concrete goals for fairness and inclusion

- Engage with social scientists, humanists, and other relevant experts for your product to understand and account for various perspectives.
- Consider how the technology and its development over time will impact different use cases:
  - Whose views are represented?
  - What types of data are represented?
  - What biases, negative experiences, or discriminatory outcomes might occur?
- Set goals for your system to work fairly across anticipated use cases: for example, in X different languages, or to Y different age groups. Monitor these goals over time and expand as appropriate.
- Update your training and testing data frequently based on who uses your technology and how they use it.

Fairness: Recommended practices

- It is important to identify whether or not machine learning can help provide an adequate solution to the specific problem at hand.

- If it can, just as there is **no single "correct" model** for all tasks, there is **no** single technique that ensures fairness **in every situation.**

Use representative datasets to train and test your model

- Assess fairness in your datasets, which includes identifying representation and corresponding limitations, and prejudicial or discriminatory correlations between features, labels, and groups.

- Public training datasets will often need to be augmented to better reflect real-world frequencies of people, events, and attributes that your system will be making predictions about.

- Understand the various perspectives, experiences, and goals of the people annotating the data. Account for human variability, including accessibility, muscle memory, and biases in annotation, e.g., by using a standard set of questions with known answers.

Fairness: Recommended practices

- It is important to identify whether or not machine learning can help provide an adequate solution to the specific problem at hand.

- If it can, just as there is **no single “correct” model** for all tasks, there is **no** single technique that ensures fairness **in every situation.**

Check the system for unfair biases

- While designing metrics to train and evaluate your system, also include metrics to examine performance across different subgroups.
  - For example, false positive rate and false negative rate per subgroup can help to understand which groups experience disproportionately worse or better performance.

- Create a test set that stress-tests the system on difficult cases.
  - How well your system is doing on examples that can be particularly hurtful or problematic each time you update your system.

- Consider the effects of biases created by decisions made by the system previously, and the feedback loops this may create.

Fairness: Recommended practices

- It is important to identify whether or not machine learning can help provide an adequate solution to the specific problem at hand.

- If it can, just as there is **no single "correct" model** for all tasks, there is **no** single technique that ensures fairness **in every situation.**

Analyse performance

- Analyse the selected metrics as a whole.
    - For example, a system's false positive rate may vary across different subgroups in your data, and improvements in one metric may adversely affect another.

- Evaluate user experience in real-world scenarios across a broad spectrum of users and use cases.

- Models rarely operate with 100% perfection when applied to real, live data. When an issue occurs in a live product, consider whether it aligns with any existing societal disadvantage, and how it will be impacted by both short- and long-term solutions.

Responsible AI practices

Interpretability

- Interpretability is crucial to being able to question, understand, and trust AI systems.
- Reflects domain knowledge and societal values.
- Provides scientists and engineers with better means of designing, developing, and debugging models.
- Helps to ensure that AI systems are working as intended.
- Understanding and testing AI systems also offers new challenges compared to traditional software.

Responsible AI practices

Interpretability

- Interpretability is crucial to being able to question, understand, and trust AI systems.
- Reflects domain knowledge and societal values.
- Provides scientists and engineers with better means of designing, developing, and debugging models.
- Helps to ensure that AI systems are working as intended.
- Understanding and testing AI systems also offers new challenges compared to traditional software.

Complex AI models, such as deep neural networks, include millions of parameters and mathematical operations, and it is much harder to pinpoint one specific bug that leads to a faulty decision.

Traditional software is essentially a series of if-then rules, and interpreting and debugging performance largely consists of chasing a problem down a garden of forking paths. While that can be gnarly, a human can generally track the path taken through the code, and understand a given result.

Responsible AI practices

Interpretability

- Interpretability is crucial to being able to question, understand, and trust AI systems.
- Reflects domain knowledge and societal values.
- Provides scientists and engineers with better means of designing, developing, and debugging models.
- Helps to ensure that AI systems are working as intended.
- Understanding and testing AI systems also offers new challenges compared to traditional software.

Complex AI models, such as deep neural networks, include millions of parameters and mathematical operations, and it is much harder to pinpoint one specific bug that leads to a faulty decision.

values can be traced to the training data or model

Traditional software is essentially a series of if-then rules, and interpreting and debugging performance largely consists of chasing a problem down a garden of forking paths. While that can be gnarly, a human can generally track the path taken through the code, and understand a given result.

“magic numbers”, magic thresholds or now-forgotten rules, personal intuition

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Plan out your options to pursue interpretability

- Pursuing interpretability can happen before, during and after designing and training your model.
- What degree of interpretability is really needed?
  - Work closely with relevant domain experts for your model (e.g., healthcare, retail, etc.) to identify what interpretability features are needed, and why.
- Can you analyse your training/testing data?
  - For example, if you are working with private data, you may not have access to investigate your input data.
- Can you change your training/testing data, for example, gather more training data for certain subsets (e.g., parts/slices of the feature space), or gather test data for categories of interest?
- Can you design a new model or are you constrained to an already-trained model?
- Are you providing too much transparency, potentially opening up vectors for abuse?
- What are your post-train interpretability options? Will you have access to the internals of the model (e.g., black box vs. white box)?

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Treat interpretability as a core part of the user experience

- Iterate with users in the development cycle to test and refine your assumptions about user needs and goals.
- Design the UX so that users build useful mental models of the AI system.
  - If not given clear and compelling information, users may make up their own theories about how an AI system works, which can negatively affect how they try to use the system.
- Where possible, make it easy for users to do their own sensitivity analysis: **empower** them to test how different inputs affect the model output.

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Design the model to be interpretable

- Use the **smallest** set of inputs necessary for your performance goals to make it clearer what factors are affecting the model.
- Use the **simplest** model that meets your performance goals.
- Learn causal relationships not correlations when possible.
  - For example, use height not age to predict if a kid is safe to ride a roller coaster.
- Craft the training objective to match your true goal.
  - For example, train for the acceptable probability of false alarms, not accuracy.
- Constrain your model to produce input-output relationships that reflect domain expert knowledge.
  - For example, a coffee shop should be more likely to be recommended if it's closer to the user, if everything else about it is the same.

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Choose metrics to reflect the end-goal and the end-task

- The metrics you consider must address the particular benefits and risks of your specific context.
  - For example, a fire alarm system would need to have high recall, even if that means the occasional false alarm.
- Analyse the model’s sensitivity to different inputs, for different subsets of examples.

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Communicate explanations to model users

- Provide explanations that are understandable and appropriate for the user.
  - For example, technical details may be appropriate for industry practitioners and academia, while general users may find UI prompts, user-friendly summary descriptions or visualizations more useful.

- **Explanations should be informed by a careful consideration of philosophical, psychological, computer science (including HCI), legal and ethical considerations about what counts as a good explanation in different contexts.**

- Identify if and where **explanations may not be appropriate**
  - For example, where explanations could result in more confusion for general users, nefarious actors could take advantage of the explanation for system or user abuse, or explanations may reveal proprietary information.

- Prioritize explanations that suggest clear actions a user can take to correct inaccurate predictions going forward.

Responsible AI practices

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Communicate explanations to model users (2)

- Consider alternatives if explanations are requested by a certain user but cannot or should not be provided, or if it's not possible to provide a clear, sound explanation.
- Don’t imply that explanations mean causation unless they do.
- Recognize human psychology and limitations (e.g., confirmation bias, cognitive fatigue).
- Explanations can come in many forms → use best practices from HCI and visualization.
- Be mindful of the **limitations** of explanations.
  - For example, local explanations may not generalize broadly, and may provide conflicting explanations of two visually-similar examples.

Responsible AI practices

Interpretability: Recommended practices

- An AI system is best understood by the underlying training data and training process, as well as the resulting AI model.

Test, Test, Test

- Conduct rigorous unit tests to test each component of the system in isolation.
- Proactively detect input drift by testing the statistics of the inputs to the AI system to make sure they are not changing in unexpected ways.
- Use a gold standard dataset to test the system and ensure that it continues to behave as expected.
  - **Update this test set regularly in line with changing users and use cases!!**
- Conduct integration tests: understand how the AI system interacts with other systems and what, if any, feedback loops are created.
  - For example, recommending a news story because it's popular can make that news story more popular, causing it to be recommended more.

Responsible AI practices

Privacy

- Sometimes the training data, input data, or both can be quite sensitive.
  - We don’t want it to go public!

- Although there may be enormous **benefits** to building a model that operates on sensitive data, it is essential to consider the **potential privacy implications** in using sensitive data.

- Not only respecting the legal and **regulatory requirements**, but also considering **social norms** and typical individual expectations.
  - What safeguards need to be put in place to ensure the privacy of individuals considering that ML models may remember or reveal aspects of the data they have been exposed to?
  - What steps are needed to ensure users have adequate transparency and control of their data?

Responsible AI practices

Privacy

e.g., a cancer detector trained on a dataset of biopsy images and deployed on individual patient scans

- Sometimes the training data, input data, or both can be quite sensitive.
  - We don't want it to go public!
- Although there may be enormous **benefits** to building a model that operates on sensitive data, it is essential to consider the **potential privacy implications** in using sensitive data.
- Not only respecting the legal and **regulatory requirements**, but also considering **social norms** and typical individual expectations.
  - What safeguards need to be put in place to ensure the privacy of individuals considering that ML models may remember or reveal aspects of the data they have been exposed to?
  - What steps are needed to ensure users have adequate transparency and control of their data?

Privacy: Recommended practices

- There is no single correct approach to ML privacy protection across all scenarios.
- Balance privacy and utility.
- Privacy needs to be clearly defined.

Collect and handle data responsibly

- Identify whether your ML model can be trained **without** the use of sensitive data.

- For example, by using non-sensitive data collection or an existing public data source.
  - Strive to minimize the use of sensitive data.

- Handle any sensitive data with care.
  - For example, comply with required laws and standards, provide users with clear notice and give them any necessary controls over data use.

- Anonymize and aggregate incoming data using best practice data-scrubbing pipelines.
  - For example, consider removing personally identifiable information and outlier or metadata values that might allow de-anonymisation.

Responsible AI practices

Privacy: Recommended practices

- There is no single correct approach to ML privacy protection across all scenarios.
- Balance privacy and utility.
- Privacy needs to be clearly defined.

Leverage on-device processing where appropriate

- If the goal is to learn statistics of individual interactions, consider collecting only statistics that have been computed locally, on-device, rather than raw interaction data, which can include sensitive information.
- When feasible, apply aggregation, randomization, and scrubbing operations on-device.

Privacy: Recommended practices

- There is no single correct approach to ML privacy protection across all scenarios.
- Balance privacy and utility.
- Privacy needs to be clearly defined.

Appropriately safeguard the privacy of ML models

- It is crucial to consider the privacy impact of how the models were constructed and may be accessed.
- Estimate whether your model is unintentionally memorizing or exposing sensitive data.
- Experiment with parameters for data minimization (e.g., aggregation, outlier thresholds, and randomization factors) to understand trade-offs and identify optimal settings for your model.
- Train ML models using techniques that establish mathematical guarantees for privacy.
- Follow best-practice processes established for cryptographic and security-critical software.

Responsible AI practices

Security

- Safety and security entails ensuring AI systems behave as intended, regardless of how attackers try to interfere.

- It is essential to consider and address the security of an AI system before it is widely relied upon in safety-critical applications.

- **Challenges**.
  - It is hard to predict all scenarios ahead of time, especially when ML is applied to problems that are difficult for humans to solve.
  - It is hard to build systems that provide both the necessary restrictions for security as well as the necessary flexibility to generate creative solutions or adapt to unusual inputs.
  - New ways of attacking constantly appear. Hard to keep up.

Responsible AI practices

Security: Recommended practices

- Security research in ML spans a wide range of threats.

- Developers should think about whether their system is likely to come under attack, consider the likely consequences of a successful attack and in most cases should simply not build systems where such attacks are likely to have significant negative impact.

Identify potential threats to the system

- Consider whether anyone would have an incentive to make the system misbehave.

- Identify what unintended consequences would result from the system making a mistake, and assess the likelihood and severity of these consequences.

- Build a rigorous threat model to understand all possible attack vectors.
  - For example, a system that would allow an attacker to change the input to the ML model may be much more vulnerable than a system that processes metadata collected by the server, like timestamps of actions the user took, since it is much harder for a user to intentionally modify input features collected without their direct participation.

Summary

- AI is not a silver bullet, but it could help tackle some of the world's most challenging social problems.
- The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education.
- AI tools and techniques can be misused so principles for their use must be established.
- Fairness
  - AI solutions can unintentionally harm the very people they are supposed to help.
  - Systems have the potential to be **fairer** and **more inclusive** at a broader scale than decision-making processes based on **ad hoc rules or human judgments.**
- Interpretability
  - Crucial to being able to question, understand, and trust AI systems.
  - Reflects domain knowledge and societal values.
- Privacy
  - Although there may be enormous **benefits** to building a model that operates on sensitive data, it is essential to consider the **potential privacy implications** in using sensitive data.
- Security
  - It is essential to consider and address the security of an AI system before it is widely relied upon in safety-critical applications.

Thanks!

Questions?

ACM Summer School on User Modeling and Personalization in Urban Computing:

AI for social good

Dr Antonela Tommasel

ISISTAN, CONICET-UNICEN, Argentina

CONICET

ISISTAN