Are PETs and Algorithmic Accountability at loggerheads?

Anders Dalskov & Kris Shrishak

HotPETs 2023

Partisia

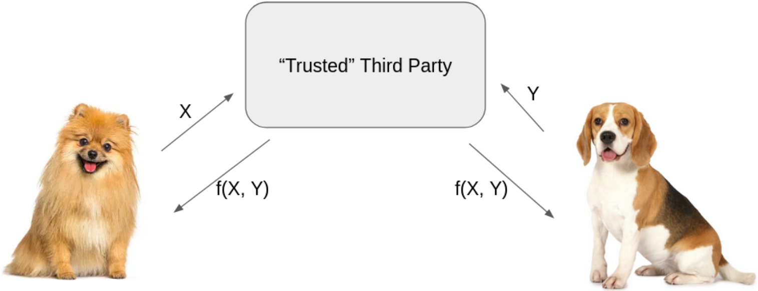
Irish Council for Civil Liberties

MPC

Secure Multiparty Computation (MPC) is the PET of choice for this talk

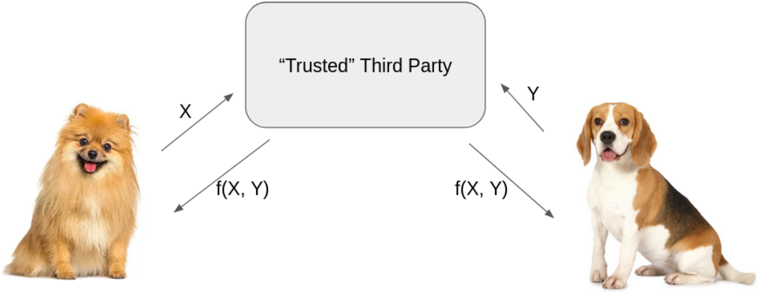
MPC

Secure Multiparty Computation (MPC) is the PET of choice for this talk



MPC

Secure Multiparty Computation (MPC) is the PET of choice for this talk



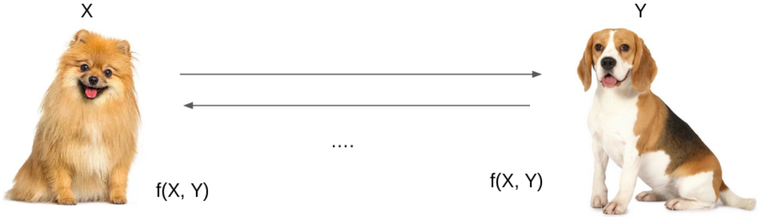
Clearly private and correct, provided the "trusted" third party is actually trusted

MPC

MPC is a type of interactive protocol between participants

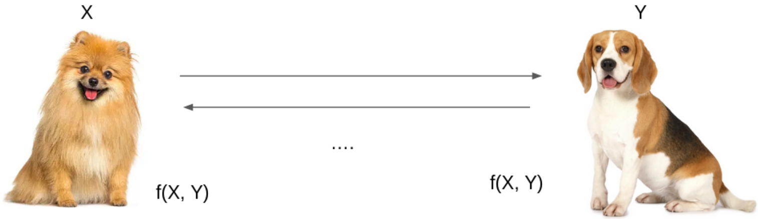
MPC

MPC is a type of interactive protocol between participants



MPC

MPC is a type of interactive protocol between participants



Provides the same guarantees as if there was a trusted party :)

MPC

MPC comes in a wide variety of flavours:

MPC

MPC comes in a wide variety of flavours:

1. Honest majority, honest super-majority, dishonest majority, ...

MPC

MPC comes in a wide variety of flavours:

1. Honest majority, honest super-majority, dishonest majority, ...
2. Malicious adversary, covert adversary, mixed adversary, semi-honest adversary, ...

MPC

MPC comes in a wide variety of flavours:

1. Honest majority, honest super-majority, dishonest majority, ...
2. Malicious adversary, covert adversary, mixed adversary, semi-honest adversary, ...
3. Computational security, statistical security, everlasting security, perfect security, ...

MPC

MPC comes in a wide variety of flavours:

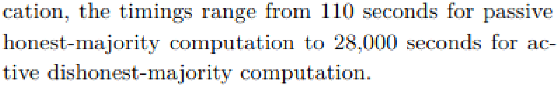
1. Honest majority, honest super-majority, dishonest majority, ...
2. Malicious adversary, covert adversary, mixed adversary, semi-honest adversary, ...
3. Computational security, statistical security, everlasting security, perfect security, ...
4. Security with abort, fairness, guaranteed output delivery, identifiable abort, ...

MPC

MPC comes in a wide variety of flavours:

1. Honest majority, honest super-majority, dishonest majority, ...
2. Malicious adversary, covert adversary, mixed adversary, semi-honest adversary, ...
3. Computational security, statistical security, everlasting security, perfect security, ...
4. Security with abort, fairness, guaranteed output delivery, identifiable abort, ...

Practicality depends on the flavour. e.g.,



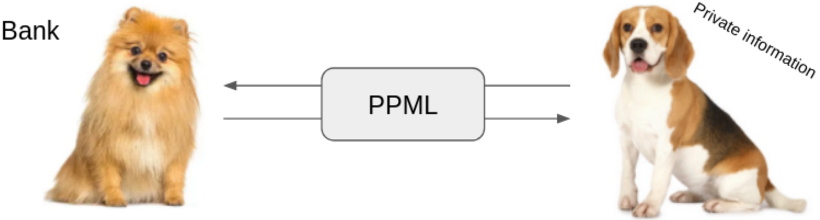
cation, the timings range from 110 seconds for passive honest-majority computation to 28,000 seconds for active dishonest-majority computation.

MPC and ML = PPML

Let's consider an realistic application of PPML

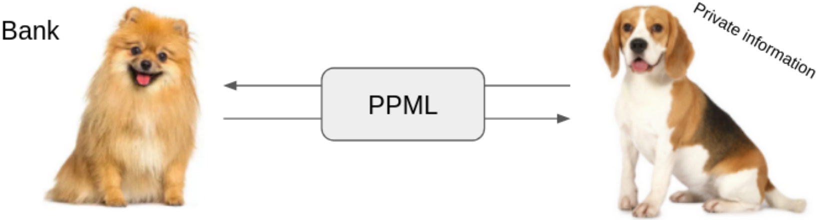
MPC and ML = PPML

Let's consider an realistic application of PPML



MPC and ML = PPML

Let's consider an realistic application of PPML



The customer doesn't trust the bank with all its private information, so it resorts to PPML.

Algorithmic Accountability (AA)

Algorithmic Accountability is about an obligation to report, explain or justify the outputs of an algorithm[1]

[1] www.fatml.org/resources/principles-for-accountable-algorithms

Algorithmic Accountability (AA)

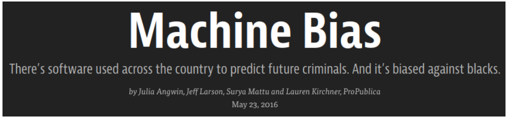
Algorithmic Accountability is about an obligation to report, explain or justify the outputs of an algorithm[1]

"*The algorithm did it*" is not an acceptable response when the "AI" misbehaves

[1] www.fatml.org/resources/principles-for-accountable-algorithms

Algorithmic Accountability (AA)

Algorithmic Accountability (AA)



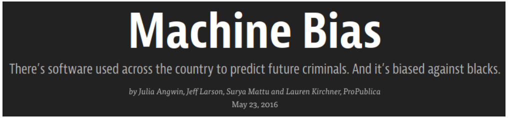
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Algorithmic Accountability (AA)

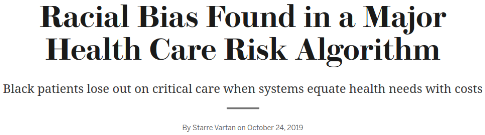


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

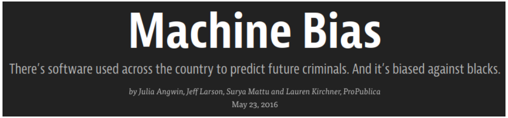


Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Algorithmic Accountability (AA)

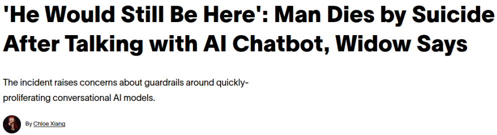


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

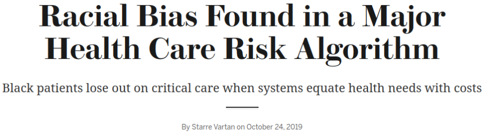
May 23, 2016



'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

By Chloe Xiang

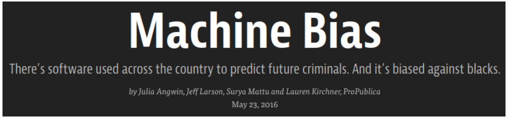


Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Algorithmic Accountability (AA)

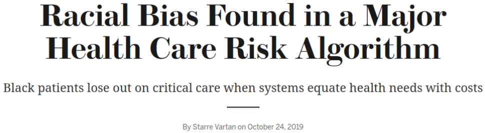


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

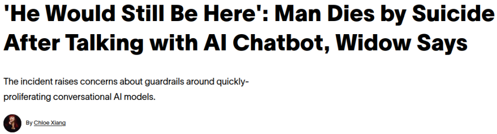
May 23, 2016



Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

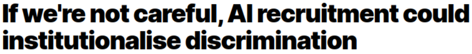
By Starre Vartan on October 24, 2019



'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

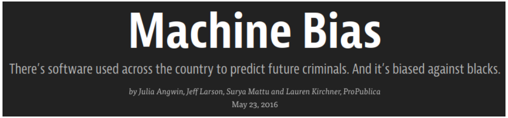
The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

By Chloe Xiang



If we're not careful, AI recruitment could institutionalise discrimination

Algorithmic Accountability (AA)

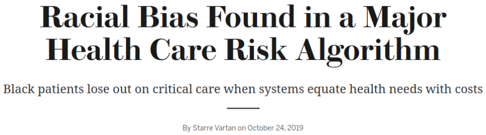


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

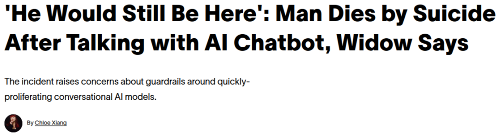
May 23, 2016



Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

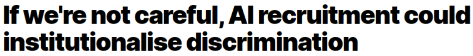
By Starre Vartan on October 24, 2019



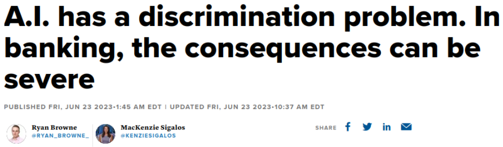
'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

By Chloe Xiang



If we're not careful, AI recruitment could institutionalise discrimination



A.I. has a discrimination problem. In banking, the consequences can be severe

PUBLISHED FRI, JUN 23 2023•1:45 AM EDT | UPDATED FRI, JUN 23 2023•10:37 AM EDT

Ryan Browne @RYAN_BROWNE_

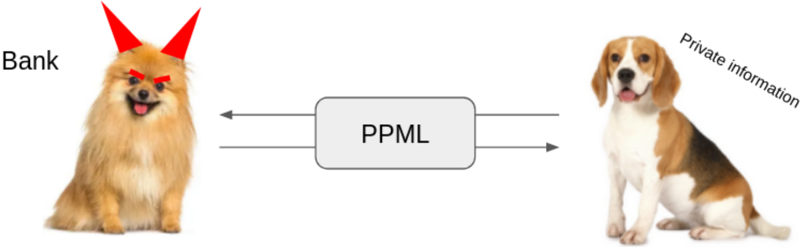
MacKenzie Sigalos @KENZIESIGALOS

AA?

Certainly, MPC helps if we wish to protect privacy.

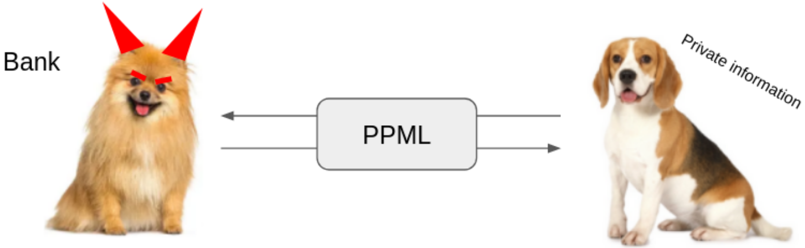
AA?

Certainly, MPC helps if we wish to protect privacy.



AA?

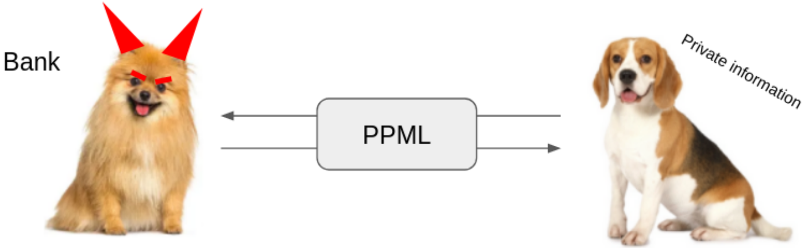
Certainly, MPC helps if we wish to protect privacy.



So we're good, right?

AA?

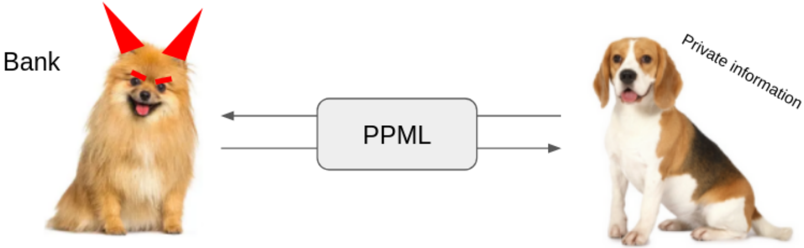
Certainly, MPC helps if we wish to protect privacy.



So we're good, right? Privacy yes.

AA?

Certainly, MPC helps if we wish to protect privacy.



Bank

PPML

Private information

So we’re good, right? Privacy yes.
But what if the goal of the adversary is to discriminate against the customer?

Inputs

Security definitions for MPC don't care if the adversary uses M_{fair} or M_{unfair} as its input

Inputs

Security definitions for MPC don’t care if the adversary uses M_{fair} or M_{unfair} as its input

That means it’s the responsibility of the *function* (not the protocol) to ensure accountability for the output.

Inputs

Security definitions for MPC don't care if the adversary uses M_{fair} or M_{unfair} as its input

That means it's the responsibility of the *function* (not the protocol) to ensure accountability for the output.

But that seems hard—instead of a secure computation that does

$$y = \mathsf{Eval}([M], [x])$$

Inputs

Security definitions for MPC don't care if the adversary uses M_{fair} or M_{unfair} as its input

That means it's the responsibility of the *function* (not the protocol) to ensure accountability for the output.

But that seems hard—instead of a secure computation that does

$$y = \mathsf{Eval}([M], [x])$$

We now have to perform a secure computation that does

$$y = \mathsf{EvalButOnlyIfModelIsGoodAccordingToGoodThatAgreesWithThreatModel}([M], [x])$$

Prior research

The apparant clash between privacy and AA have been observed before

- *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML* (IEEE Access 2022)
- *Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects* (Arxiv 2023)

Prior research

The apparant clash between privacy and AA have been observed before

- *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML* (IEEE Access 2022)
- *Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects* (Arxiv 2023)

Others attack the issue explicity, e.g.,

- *Private and Reliable Neural Network Inference* (CCS 2022)

Prior research

The apparant clash between privacy and AA have been observed before

- *An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML* (IEEE Access 2022)
- *Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects* (Arxiv 2023)

Others attack the issue explicity, e.g.,

- *Private and Reliable Neural Network Inference* (CCS 2022)

And others attack this issue, but from the “wrong” direction

- *Planting Undetectable Backdoors in Machine Learning Models* (FOCS 2022)

Up for discussion

Thus we arrive at a (non-exhaustive) list of questions

Up for discussion

Thus we arrive at a (non-exhaustive) list of questions

- First of all, is this even a problem?

Up for discussion

Thus we arrive at a (non-exhaustive) list of questions

- First of all, is this even a problem?
- Is this an MPC issue, or does it apply to other PETs? Are some PETs “immune” to this problem?

Up for discussion

Thus we arrive at a (non-exhaustive) list of questions

- First of all, is this even a problem?
- Is this an MPC issue, or does it apply to other PETs? Are some PETs “immune” to this problem?
- Do PETs, specifically in the context of PPML, fail if they cannot also facilitate AA?

Up for discussion

Thus we arrive at a (non-exhaustive) list of questions

- First of all, is this even a problem?
- Is this an MPC issue, or does it apply to other PETs? Are some PETs “immune” to this problem?
- Do PETs, specifically in the context of PPML, fail if they cannot also facilitate AA?
- Is this a PPML only issue?