The Security & Privacy Implications of AI and machine learning

Peter Eckersley
Chief Computer Scientist
Electronic Frontier Foundation
pde@eff.org
What do we mean by AI?
“Narrow” AI (NAI) – mostly machine learning algorithms, often deep / convolutional neural nets

“General” AI (AGI) – the stuff of science fiction, may or may not happen anytime soon
AI is getting a lot of hype right now

(partly NAI progress, partly AGI speculation)
Why is everyone worked up about NAI?
Privacy refers to the natural person entitled to a private life of tranquility and personal information secret by law protected against unlawful intrusion, knows the collection, use and disclosure of the right kind of personality, but also the rights of others can subject yourself to what extent intervention of private life, whether their privacy is open to others and the scope and extent of the public and so have the right to decide. Privacy is a basic personality right.

Two translations of Baidu’s privacy encyclopedia entry:

SYSTRAN: The right of privacy is the personal life peaceful and personal information secret that to refer to the natural person enjoys is protected according to law, was not invaded, learned about, collected and uses and public one personality power by other people illegally. Moreover the subject of object can involve own private life to other people in what degree, to own privacy whether had the decision-

Neural Machine Translation: Privacy refers to the natural person entitled to a private life of tranquility and personal information secret by law protected against unlawful intrusion, knows the collection, use and disclosure of the right kind of personality, but also the rights of others can subject yourself to what extent intervention Of private life , whether their privacy is open to others and the scope and extent of the public and so
Two translations of Baidu’s privacy encyclopedia entry:

SYSTRAN: The right of privacy is the personal life peaceful and personal information secret that to refer to the natural person enjoys is protected according to law, was not invaded, learned about, collected and uses and public one personality power by other people illegally. Moreover the subject of object can involve own private life to other people in what degree, to own privacy whether had the decision-

Neural Machine Translation: Privacy refers to the natural person entitled to a private life of tranquility and personal information secret by law protected against unlawful intrusion, knows the collection, use and disclosure of the right kind of personality, but also the rights of others can subject yourself to what extent intervention Of private life, whether their privacy is open to others and the scope and extent of the public and so
<table>
<thead>
<tr>
<th>(a) CelebA</th>
<th>17 bits/px</th>
<th>(b) JPEG2000</th>
<th>29x compression</th>
<th>0.580</th>
<th>20.49</th>
<th>0.588</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c) JPEG</td>
<td>24x compression</td>
<td></td>
<td></td>
<td>0.718</td>
<td>20.33</td>
<td>0.590</td>
</tr>
<tr>
<td>(d) NCode(16)</td>
<td>44x compression</td>
<td></td>
<td></td>
<td>0.390</td>
<td>20.06</td>
<td>0.712</td>
</tr>
<tr>
<td>(e) NCode(4)</td>
<td>173x compression</td>
<td></td>
<td></td>
<td>0.098</td>
<td>19.84</td>
<td>0.704</td>
</tr>
<tr>
<td>(f) NCode(2)</td>
<td>347x compression</td>
<td></td>
<td></td>
<td>0.049</td>
<td>16.20</td>
<td>0.573</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a) UT Zappos50k</th>
<th>11 bits/px</th>
<th>(b) JPEG2000</th>
<th>16x compression</th>
<th>0.693</th>
<th>19.64</th>
<th>0.705</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c) JPEG</td>
<td>15x compression</td>
<td></td>
<td></td>
<td>0.750</td>
<td>19.90</td>
<td>0.707</td>
</tr>
<tr>
<td>(d) NCode(16)</td>
<td>28x compression</td>
<td></td>
<td></td>
<td>0.391</td>
<td>18.82</td>
<td>0.732</td>
</tr>
<tr>
<td>(e) NCode(4)</td>
<td>112x compression</td>
<td></td>
<td></td>
<td>0.098</td>
<td>17.14</td>
<td>0.693</td>
</tr>
<tr>
<td>(f) NCode(2)</td>
<td>224x compression</td>
<td></td>
<td></td>
<td>0.049</td>
<td>11.13</td>
<td>0.523</td>
</tr>
</tbody>
</table>

Imagenet image recognition
<table>
<thead>
<tr>
<th>Task 11: Basic Coreference</th>
<th>Task 12: Conjunction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel was in the kitchen.</td>
<td>Mary and Jeff went to the kitchen.</td>
</tr>
<tr>
<td>Then he went to the studio.</td>
<td>Then Jeff went to the park.</td>
</tr>
<tr>
<td>Sandra was in the office.</td>
<td>Where is Mary? A: kitchen</td>
</tr>
<tr>
<td>Where is Daniel? A: studio</td>
<td>Where is Jeff? A: park</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 13: Compound Coreference</th>
<th>Task 14: Time Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel and Sandra journeyed to the office.</td>
<td>In the afternoon Julie went to the park.</td>
</tr>
<tr>
<td>Then they went to the garden.</td>
<td>Yesterday Julie was at school.</td>
</tr>
<tr>
<td>Sandra and John travelled to the kitchen.</td>
<td>Julie went to the cinema this evening.</td>
</tr>
<tr>
<td>After that they moved to the hallway.</td>
<td>Where did Julie go after the park? A: cinema</td>
</tr>
<tr>
<td>Where is Daniel? A: garden</td>
<td>Where was Julie before the park? A: school</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 15: Basic Deduction</th>
<th>Task 16: Basic Induction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheep are afraid of wolves.</td>
<td>Lily is a swan.</td>
</tr>
<tr>
<td>Cats are afraid of dogs.</td>
<td>Lily is white.</td>
</tr>
<tr>
<td>Mice are afraid of cats.</td>
<td>Bernhard is green.</td>
</tr>
<tr>
<td>Gertrude is a sheep.</td>
<td>Greg is a swan.</td>
</tr>
<tr>
<td>What is Gertrude afraid of? A:wolves</td>
<td>What color is Greg? A:white</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 17: Positional Reasoning</th>
<th>Task 18: Size Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>The triangle is to the right of the blue square.</td>
<td>The football fits in the suitcase.</td>
</tr>
<tr>
<td>The red square is on top of the blue square.</td>
<td>The suitcase fits in the cupboard.</td>
</tr>
<tr>
<td>The red sphere is to the right of the blue square.</td>
<td>The box is smaller than the football.</td>
</tr>
<tr>
<td>Is the red sphere to the right of the blue square? A:yes</td>
<td>Will the box fit in the suitcase? A:yes</td>
</tr>
<tr>
<td>Is the red square to the left of the triangle? A:yes</td>
<td>Will the cupboard fit in the box? A:no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 19: Path Finding</th>
<th>Task 20: Agent’s Motivations</th>
</tr>
</thead>
<tbody>
<tr>
<td>The kitchen is north of the hallway.</td>
<td>John is hungry.</td>
</tr>
<tr>
<td>The bathroom is west of the bedroom.</td>
<td>John goes to the kitchen.</td>
</tr>
<tr>
<td>The den is east of the hallway.</td>
<td>John grabbed the apple there.</td>
</tr>
<tr>
<td>The office is south of the bedroom.</td>
<td>Daniel is hungry.</td>
</tr>
<tr>
<td>How do you go from den to kitchen? A: west, north</td>
<td>Where does Daniel go? A:kitchen</td>
</tr>
<tr>
<td>How do you go from office to bathroom? A: north, west</td>
<td>Why did John go to the kitchen? A:hungry</td>
</tr>
</tbody>
</table>
What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
1. Mass surveillance becomes effective
Until recently, learning much about people from that kind of data required the attention of an analyst.
Facebook is a pretty good indication of how rapidly that is changing
2. Privacy, bias, and algorithmic decision-making
Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson
ProPublica, Dec. 30, 2016, 4:44 p.m.
Problems:
The source data for these systems is terrible (and biased)
Problem: Omitted Variable Bias
Omitted variable bias:

True Cause is not in your data set.

True Cause is correlated with race. Race is correlated with other things (like zip code). Zip code is in the training data.
Partying and drinking predict car crashes.

Car insurance firms can’t measure that, but they can measure the time people drive.

So they penalize customers who drive late at night.

→ unfair premiums for nightshift workers
Simulating loan decisions for different groups
Drag the black threshold bars left or right to change the cut-offs for loans.
Click on different preset loan strategies.

<table>
<thead>
<tr>
<th>Loan Strategy</th>
<th>Blue Population</th>
<th>Orange Population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAX PROFIT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No constraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GROUP UNAWARE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue and orange thresholds are the same</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DEMOGRAPHIC PARITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same fractions blue / orange loans</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EQUAL OPPORTUNITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same fractions blue / orange</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **loan threshold:** 50

Okay, and how does all of this interact with privacy?
Algorithms *expect* people to make relevant information about them available through various data sources.
For instance:
denying insurance to people who use fake birthdays on Facebook (!)
3. Finding bugs
Search through program input space / execution space for exploits
DARPA's Cyber Grand Challenge: Early Highlights from the Competition
Fundamental question:

How does this affect the balance between offense & defense?
Schneier, 1998: “the attacker always wins”

“Security designers occupy what Prussian general Carl von Clausewitz calls "the position of the interior." A good security product must defend against every possible attack, even attacks that haven't been invented yet. Attackers, on the other hand, only need to find one security flaw in order to defeat the system. And they can cheat. They can collude, conspire, and wait for technology to give them additional tools. They can attack the system in ways the system designer never thought of.”
That’s probably too pessimistic

Some attackers will always get through some of your layers of defense

may be a more accurate
Interestingly, automated exploit detection seems to change that equilibrium
Scarce expensive expert auditing

→ (partially replaced by) →

Cheap fuzzer coverage
Problem:

how do we ensure that everything (IoT devices, phones, etc) get fuzzer coverage?
The Morris Internet Worm source code

This disk contains the complete source code of the Morris Internet worm program. This tiny, 99-line program brought large pieces of the Internet to a standstill on November 2nd, 1988.

The worm was the first of many intrusive programs that use the Internet to spread.

The Computer History Museum
Cryptoviral malware

Your personal files are encrypted!

Your important files encryption produced on this computer: photos, videos, documents, etc. Here is a complete list of encrypted files, and you can personally verify this.

Encryption was produced using a unique public key RSA-2048 generated for this computer. To decrypt files you need to obtain the private key.

The single copy of the private key, which will allow you to decrypt the files, located on a secret server on the Internet, the server will destroy the key after a time specified in the window. After that, nobody and never will be able to restore files...

To obtain the private key for this computer, which will automatically decrypt files, you need to pay 300 USD / 300 EUR / similar amount in another currency.

Click «next» to select the method of payment and the currency.

Any attempt to remove or damage this software will lead to the immediate destruction of the private key by server.

Private key will be destroyed on 10/9/2013 4:25 PM

Time left 95:56:35

Next >>
Some CGC participants argued for deployment via network firewalls
Problem: that’s incompatible with TLS :(

Idea: what about security agents on endpoints themselves?
Comodo 6.X Gaping Hole of DOOM

Comodo, as you may know, is a colossal pain in the posterior. It literally catches everything until you tell it not to, including standard windows services (say what?!?).

...at least, that's what happens on Comodo 5.X. In 6.X, Comodo apparently decided that catching things that were part of windows was a Bad Thing(tm). Their "fix" was... kinda lame

Anything running as SYSTEM is automatically legit under 6.X. ANYTHING. Let that sink in. Got a kernel level exploit? Good, because you can drop the kitchen sink and the contents of your garage and as long as you continue to run as SYSTEM you are golden. Yeah.

Needless to say, Comodo 6.X doesn't catch nearly as much stuff. Comodo's user base, paranoid bastards that they are, has apparently caught wind of this and lots of them haven't upgraded to 6.X. Kind of a shame, cuz this is a hole you could drive a very large wheeled freight carrying vehicle through. However, if you're lucky enough to be going against a target running 6.X, have fun!
Attacks against NAI
Adversarial Examples

Mitigation 1

include adversarial attacks in the training process, measure success against them:

https://github.com/tensorflow/cleverhans
Mitigation 2

“defensive distillation”

Generate probabilistic labelled data

Predict confidence as well as the answer

State of the art:
Makes the attack slower, but not impossible
Reward Hacking
How do we build learning systems that:

- do what you mean, not what you say

- never disable their own “off” switch

- never do something terrible, to see what happens

- function in environments that are different to their training

Paraphrasing: https://arxiv.org/abs/1606.06565
Atari 2600: Roadrunner

https://www.youtube.com/watch?v=cckZ6oAoLjo
What about AGI?

What kind of privacy issues would it raise?
DISCLAIMER: all remarks about AGI are speculative and should be used primarily for entertainment purposes
If AGI ever exists,

It is probably not going to be human-like, and may not exhibit human agency and personality

But if it did...
It might care a lot about privacy
Thoughs, memories and intentions of AGIs could be:

copied
stolen
modified

Without the permission of the AGI (or its owner)
Soul Steal

Instant

Destroy target creature with converted mana cost 3 or less. Put a token into play that's a copy of that creature.

She wasn't particularly after it, but taking his life was fun too.
If AGIs were to become a real possibility, we might want to think about how to protect them against that.
We the People

Article 1.

Section 1. All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.

Section 2. The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Number from each State shall be in proportion to its Population. But each State shall have at least one Representative.

Section 3. No Person shall be a Representative who shall not have attained to the Age of twentyfive Years, and been a Citizen of the United States for seven Years, and who shall not, when elected, be an inhabitant of that State for which he shall be chosen.

Section 4. The Term of Congress shall be two Years.

Section 5. Each House shall be the Judge of the Elections, Returns, and Votes of its own Members, and a Majority of each shall constitute a Quorum to do Business; but a smaller Number may adjourn from Day to Day, and may be authorized to compel the Attendance of absent Members, in such Manner, and under Such Penalties as each House may provide.

Section 6. The Senate of the United States shall be composed of two Senators from each State, chosen by the Legislature thereof for six Years, and each Senator shall have one Vote.

Section 7. The President of the Senate, with the Advice and Consent of the Senate, shall appoint the President pro tempore of the Senate, and also the other Officers of the Senate.

Section 8. The President of the Senate shall, at the Time of his Acceptance of the Office of President, give Notice thereof in Writing, signed by him, to the Senate. He shall deliver to the President pro tempore of the Senate when the same is next in order of succession, a Certificate of Oath or Affirmation, signed by him, that he does solemnly swear or affirm that he will faithfully execute the Office of President of the United States, unless the same shall be otherwise provided for by law.

Section 9. The President of the Senate, with the Advice and Consent of the Senate, shall appoint the President pro tempore of the Senate, and also the other Officers of the Senate.