CS 4700: Foundations of Artificial Intelligence

Spring 2020
Prof. Haym Hirsh

Lecture 37
May 11, 2020

ONLINE

This is not the semester you expected

This is not the semester you expected

You’ve gone through a lot

This is not the semester you expected

You’ve gone through a lot
It’s almost over

This is not the semester you expected

You’ve gone through a lot
It’s almost over

When you look back at this semester
aim for doing the best you could
given the cards you were given

This is not the semester you expected

You’ve gone through a lot
It’s almost over

When you look back at this semester
aim for doing the best you could
given the cards you were given

If you’re struggling, don’t do it alone
Talk to your professors, advisors, counselors, …

Today

Societal Implications of AI

Volkswagen (VW)

This article is more than 4 years old

Volkswagen under investigation over illegal software that masks emissions

California and EPA accuse VW of installing 'defeat device' software that reduces nitrogen oxide emissions while a car is undergoing official tests

Rupert Neate *in New York*
@RupertNeate
Fri 18 Sep 2015 13.17 EDT

THE WALL STREET JOURNAL.

BUSINESS | AUTOS & TRANSPORTATION | AUTOS INDUSTRY

Volkswagen Emissions Investigation Zeroes In on Two Engineers

Company investigation focuses on two men elevated after Winterkorn was made CEO

By William Boston

Updated Oct. 5, 2015 1:21 pm ET

WOLFSBURG, Germany—Two top Volkswagen engineers who found they couldn't deliver as promised a clean diesel engine for the U.S. market are at the center of a company probe into the installation of engine software designed to fool regulators, according to people

How about:

You’re working for an autonomous vehicle company and they decide to seek Defense Department contracts for autonomous {tanks, bombs, …}

How about:

You’re working for an autonomous vehicle company and they decide to seek Defense Department contracts for autonomous {tanks, bombs, …}

and you’ve been placed on the team to do it

How about:

You’re working for an autonomous vehicle company and they decide to develop technology that could be used for wide-spread surveillance – or to help find missing children

and you’ve been placed on the team to do it

How about:

You’re working for an autonomous vehicle company and they’re making inflated claims about the capabilities of their vehicle

How about:

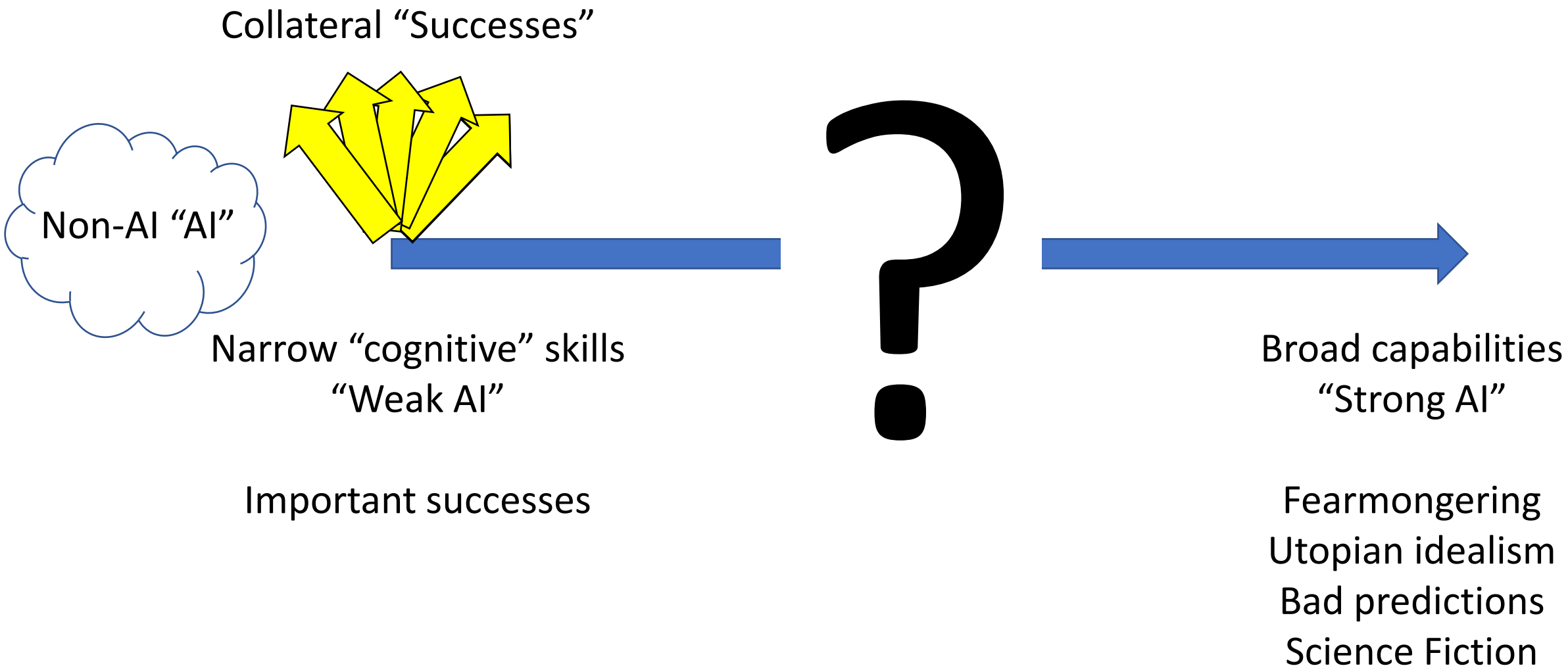
You’re working for consulting company and your client is seeking new business development opportunities and one sector they’re in involves computer vision – do you advise them about opportunities in autonomous weapons, surveillance technology, …

How about:

You’re working for a company that does natural language processing and they take on a client seeking to influence voters on platforms like Facebook and Twitter my manipulating language

and you’ve been placed on the team to do it

The AI Spectrum



The AI Spectrum

Collateral “Successes”

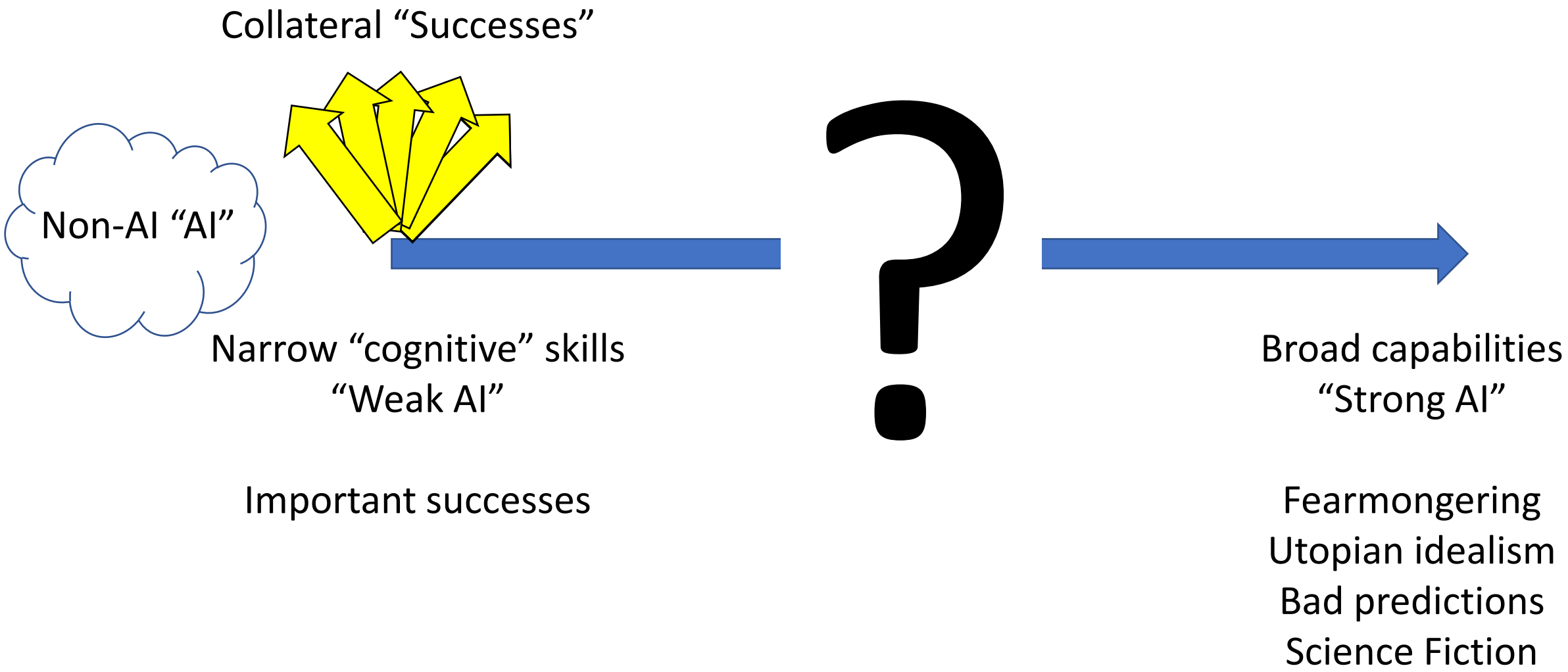
Narrow “cognitive” skills
“Weak AI”

Broad capabilities
“Strong AI”

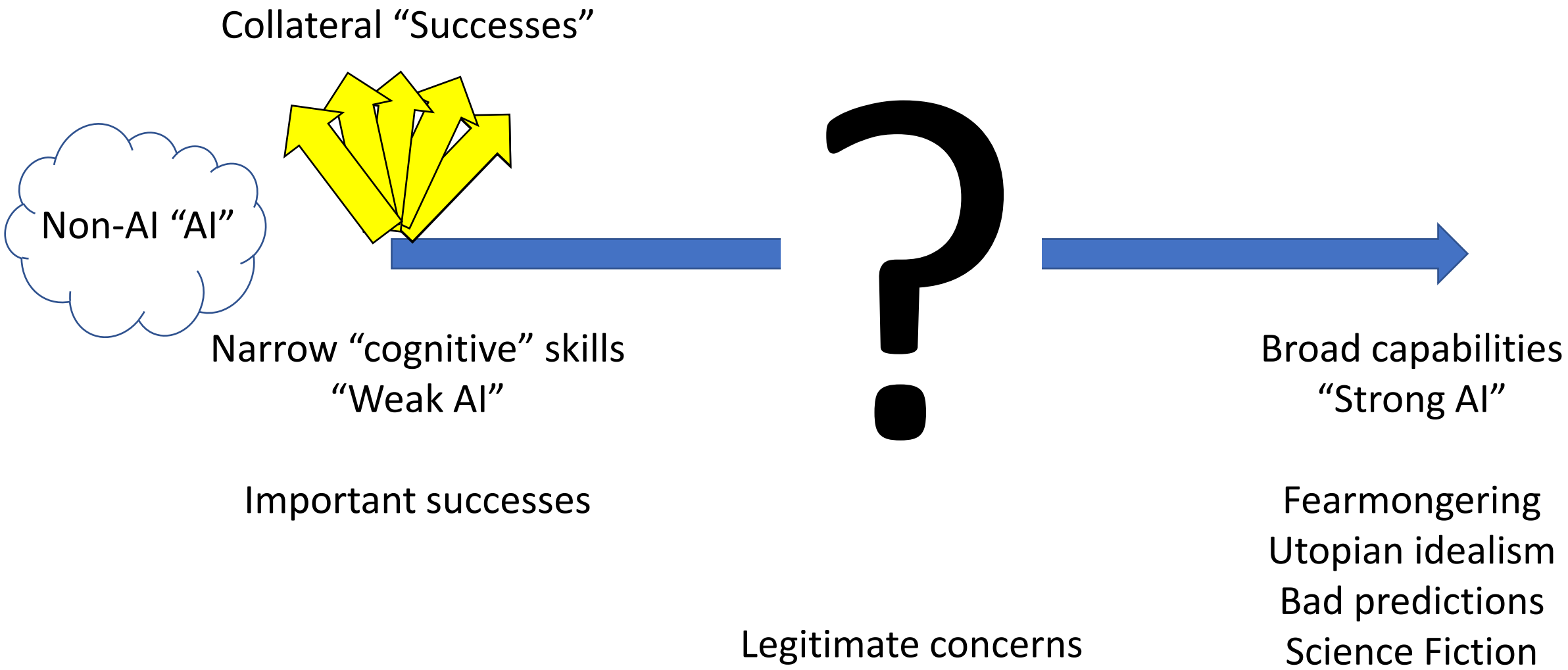
Important successes

Fearmongering
Utopian idealism
Bad predictions
Science Fiction

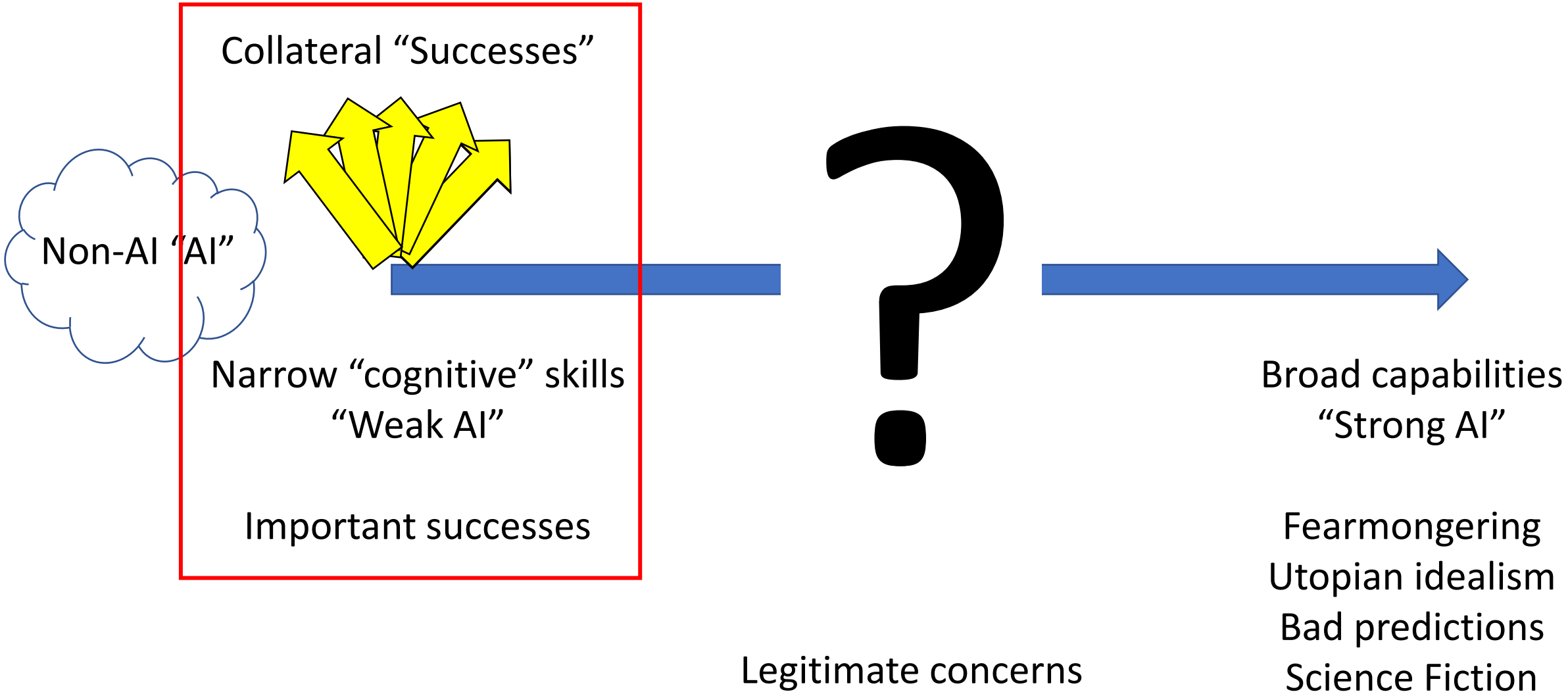
The AI Spectrum



The AI Spectrum



The AI Spectrum



Impact of Automation

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

Powered in part by AI’s collateral successes

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

- Recent and looming AI successes stand the potential to cause widespread job loss

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

- Recent and looming AI successes stand the potential to cause widespread job loss

Autonomous vehicles ⇒ 2-3 million truck drivers in US

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

- Recent and looming AI successes stand the potential to cause widespread job loss

Autonomous vehicles, robotics ⇒ 2-3M farm workers in US

Impact of Automation

- Vacation and business travel
- Taxi/transportation
- Journalism
- ….

- Recent and looming AI successes stand the potential to cause widespread job loss

Autonomous vehicles, robotics ⇒ ?M in developing world

Impact of Automation

Analogies to industrial revolution

Impact of Automation

Analogies to industrial revolution

New jobs were created to replace old jobs

Impact of Automation

Analogies to industrial revolution

New jobs were created to replace old jobs

Took six decades

Impact of Automation

We can plan for this

Impact of Automation

We can plan for this

Educational programs

Impact of Automation

We can plan for this

Educational programs

Andrew Yang: Universal income

Impact of Automation

We can plan for this

Educational programs

Andrew Yang: Universal income

Bill Gates: Tax on robots

Impact of Automation

We can plan for this

Educational programs

Andrew Yang: Universal income

Bill Gates: Tax on robots

Kai-Fu Lee: AI-leading nations bear responsibility for impact in developing world

Impact of Automation

We can plan for this

Educational programs

Andrew Yang: Universal income

Bill Gates: Tax on robots

Kai-Fu Lee: AI-leading nations bear responsibility for impact in developing world

(But it’s not clear that we are)

Impact of Automation

More generally:

Technology is increasing income divergence

What to do?

Impact of Algorithmically Mediated World

- Facebook, Google, Twitter:
  - They make more money the more you use them
  - The more you like what you see the more you use them
  - They decide what you see
  - ⇒ Give you what you want to see

Impact of Algorithmically Mediated World

- Facebook, Google, Twitter:
  - They make more money the more you use them
  - The more you like what you see the more you use them
  - They decide what you see
  - ⇒ Give you what you want to see

"Echo chambers", "Bubble filters"

Impact of Algorithmically Mediated World

- Facebook, Google, Twitter:
  - They make more money the more you use them
  - The more you like what you see the more you use them
  - They decide what you see
  - ⇒ Give you what you want to see

"Echo chambers", "Bubble filters"

- Societal impacts
  - Polarization, radicalism

Impact of Algorithmically Mediated World

- Facebook, Google, Twitter:
  - They make more money the more you use them
  - The more you like what you see the more you use them
  - They decide what you see
  - ⇒ Give you what you want to see

"Echo chambers", "Bubble filters"

- Societal impacts
  - Polarization, radicalism

Fueled by "AI"

Impact of Algorithmically Mediated World

- Facebook, Google, Twitter:
  - They make more money the more you use them
  - The more you like what you see the more you use them
  - They decide what you see
  - ⇒ Give you what you want to see

"Echo chambers", "Bubble filters"

- Societal impacts
  - Polarization, radicalism

What do you do if you're working for them?

Impact of Algorithmically Mediated World

Google changes ranking algorithm

Small businesses drop in rankings below first page

Impact of Machine Learning

Impact of Machine Learning

Values

Impact of Machine Learning

Values

- Reflecting society’s values back on itself

Technology

Google searches expose racial bias, says study of names

4 February 2013

Share

A study of Google searches has found "significant discrimination" in advert results depending on the perceived race of names searched for.

Harvard professor Latanya Sweeney said names typically associated with black people were more likely to produce ads related to criminal

Prof Sweeney said technology could be used to counteract racial intolerance

Top Stories

Johnson faces MPs as lockdown questions remain
The PM says the reopening of society and the UK economy should go ahead with "utmost caution".
11 May 2020

Iranian sailors killed in 'friendly fire' incident
1 hour ago

Wuhan in first virus cluster since end of lockdown
5 hours ago

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
    (But what about medical AI?)

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases

PRO PUBLICA

Donate

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

☐ Foster Parent(s)
☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

35. Were your brothers or sisters ever arrested, that you know of?
☐ No ☑ Yes

36. Was your wife/husband/partner ever arrested, that you know of?
☑ No ☐ Yes

37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
☑ No ☐ Yes

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
☑ No ☐ Yes

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

40. How many of your friends/acquaintances served time in jail or prison?
☐ None ☐ Few ☑ Half ☐ Most

41. How many of your friends/acquaintances are gang members?
☐ None ☑ Few ☐ Half ☐ Most

2016 WI 68

SUPREME COURT OF WISCONSIN

CASE NO.: 2015AP157-CR

COMPLETE TITLE: State of Wisconsin,
Plaintiff-Respondent,
v.
Eric L. Loomis

Ultimately, we conclude that if used properly as set forth herein, a circuit court's consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process and that the circuit court did not erroneously exercise its discretion here.

CONCURRED: ROGGENSACK, C. J. concurs (Opinion filed).
ABRAHAMSON, J. concurs (Opinion filed).

DISSENTED:

NOT PARTICIPATING:

ATTORNEYS:
For the defendant-appellant, there were briefs by *Michael D. Rosenberg* and *Community Justice, Inc.*, Madison, and oral

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

1 Introduction

Research on word embeddings has drawn significant interest in machine learning and natural language processing. There have been hundreds of papers written about word embeddings and their applications, from Web search [22] to parsing Curriculum Vitae [12]. However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.

A word embedding, trained on word co-occurrence in text corpora, represents each word (or common phrase) w as a d-dimensional *word vector* $\vec{w} \in \mathbb{R}^d$. It serves as a dictionary of sorts for computer programs that would like to use word meaning. First, words with similar semantic meanings tend to have vectors that are close together. Second, the vector differences between words in embeddings have been shown to represent relationships between words [27, 21]. For example given an analogy puzzle, "man is to king as woman is to x" (denoted as *man:king :: woman:x*), simple arithmetic of the embedding vectors finds that x=*queen* is the best answer because $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$. Similarly, x=*Japan* is returned for *Paris:France :: Tokyo:x*. It is surprising that a simple vector arithmetic can simultaneously capture a variety of relationships. It has also excited practitioners because such a tool could be useful across applications involving natural language. Indeed, they are being studied and used in a variety of downstream applications (e.g., document ranking [22], sentiment analysis [14], and question retrieval [17]).

However, the embeddings also pinpoint sexism implicit in text. For instance, it is also the case that:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

In other words, the same system that solved the above reasonable analogies will offensively answer "man is to computer programmer as woman is to x" with x=*homemaker*. Similarly, it outputs that a

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

She is a doctor
He is a babysitter

34/5000

O bir doktor
O bir bebek bakıcısı

O bir doktor
O bir bebek bakıcısı

33/5000

He is a doctor
She's a babysitter

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines
  - Gender bias in language

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines
  - Gender bias in language

  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines
  - Gender bias in language

  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

Creates feedback loops

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines
  - Gender bias in language

  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

We're growing our understanding of best practices to avoid this

Impact of Machine Learning

Values

- Reflecting society's values back on itself
  - Ad selection reflecting racial biases
  - Racial bias in sentencing guidelines
  - Gender bias in language

  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

How to regulate it? (Policy implications: Solon Barocas)

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

Was in use and undetected

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects

Was in use and undetected

We’re growing our understanding of best practices to avoid this

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability

2014

(a) Siberian husky (b) Eskimo dog

Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge.

2016

"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: *if the users do not trust a model or a prediction, they will not use it.* It is important to differentiate between two different (but related) definitions of trust: (1) *trusting a prediction*, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) *trusting a model*, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative instances with explanations to address the "trusting the model" problem, via submodular optimization.
- Comprehensive evaluation with simulated and human subjects, where we measure the impact of explanations on trust and associated tasks. In our experiments, non-experts using LIME are able to pick which classifier from a pair generalizes better in the real world. Further, they are able to greatly improve an untrustworthy classifier trained on 20 newsgroups, by doing feature engineering using LIME. We also show how understanding the predictions of a neural network on images helps practitioners know when and why they should not trust a model.

2. THE CASE FOR EXPLANATIONS

By "explaining a prediction", we mean presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction. We argue that explaining predictions is an important aspect in

KDD 2016 San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08... $15.00
DOI: http://dx.doi.org/10.1145/2939672.2939778

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability
  - Making sure classifier is correct

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

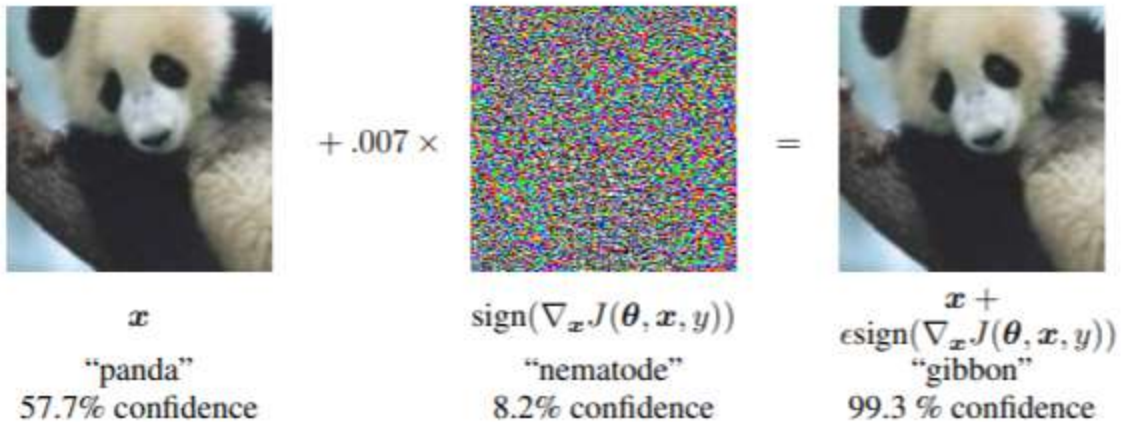
ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

1 INTRODUCTION

Szegedy et al. (2014b) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*. That is, these machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots in our training algorithms.

The cause of these adversarial examples was a mystery, and speculative explanations have suggested


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

This paper appears at CVPR 2018

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt[*1], Ivan Evtimov[*2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

Abstract

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. We propose a general attack algorithm, Robust Physical Perturbations (RP_2), to generate robust visual adversarial perturbations under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP_2 achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Due to the current lack of a standardized testing method, we propose a two-stage evaluation methodology for robust physical adversarial examples consisting of lab and field tests. Using this methodology, we evaluate the efficacy of physical adversarial manipulations on real objects. With a perturbation in the form of only black and white stickers, we attack a real stop sign, causing targeted misclassification in 100% of the images obtained in lab settings, and in 84.8% of the captured video frames obtained on a moving vehicle (field test) for the target classifier.

1. Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art, and sometimes human-competitive, performance on many computer vision tasks [11, 14, 36]. Based on these successes, they are increasingly being used as part of control pipelines in physical systems such as cars [8, 17], UAVs [4, 24], and robots [40]. Recent work, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [5, 9, 10, 15, 16, 22, 25, 29, 30, 35]. These carefully crafted modifications to the (visual) input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

This threat has gained recent attention, and work in computer vision has made great progress in understanding the space of adversarial examples, beginning in the digital domain (*e.g.* by modifying images corresponding to a scene) [9, 22, 25, 35], and more recently in the physical domain [1, 2, 13, 32]. Along similar lines, our work contributes to the understanding of adversarial examples when perturbations are physically added to the *objects themselves*. We choose road sign classification as our target domain for several reasons: (1) The relative visual simplicity of road signs make it challenging to hide perturbations. (2) Road signs exist in a noisy unconstrained environment with changing physical conditions such as the distance and angle of the viewing camera, implying that physical adversarial perturbations should be robust against considerable environmental instability. (3) Road signs play an important role in transportation safety. (4) A reasonable threat model for transportation is that an attacker might not have control over a vehicle's systems, but is able to modify the objects in the physical world that a vehicle might depend on to make crucial safety decisions.

The main challenge with generating robust physical perturbations is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms [19]. For our chosen application area, the most dynamic environmental change is the distance and angle of

[*]These authors contributed equally.


STOP

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability
  - Making sure classifier is correct
  - Adversarial examples

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability
  - Making sure classifier is correct
  - Adversarial examples
  - Adversarial training

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability
  - Making sure classifier is correct
  - Adversarial examples
  - Adversarial training (spam detection, leaking data)

Impact of Machine Learning

(Lack of) Transparency

- Data bias
  - Bias in pedestrian detection for light-skinned vs dark-skinned individuals
  - Bias in speech recognition for non-Silicon Valley dialects
- Opaque predictions vs interpretability
  - Making sure classifier is correct
  - Adversarial examples
  - Adversarial training (spam detection, leaking data)
  - Black box vs white box

Impact of Machine Learning

(Lack of) Transparency

- Ongoing directions
  - Visualizing classifiers
    - Visualizing hidden units
    - Important pixels in images
    - Examples that activate different sets of neurons
  - Certifying classifiers
  - Interpretable classifiers
    - Learn in interpretable representation
    - Learn interpretable approximations
    - Asking “what if” questions

Dual Use Challenges

AI technology can lead to societally valuable outcomes

Dual Use Challenges

AI technology can lead to societally valuable outcomes

The same AI technology can lead to societally questionable outcomes

Dual Use Challenges

AI technology can lead to societally valuable outcomes

The same AI technology can lead to societally questionable outcomes

Autonomous weapons

Surveillance state

Bugs

All software has bugs

Bugs

All software has bugs

AI has distinct bugs:

Formulating problems as learning from data

Formulating problems as optimization problems

Bugs

All software has bugs

AI has distinct bugs:

Formulating problems as learning from data

Formulating problems as optimization problems

“Value Alignment Problem”

Data

- Centralization puts sensitive data at risk
- Reverse engineering classifiers to uncover data
- Identifying outliers

- Solution approaches:
  - Federate learning
  - Privacy preserving learning
  - On device processing
  - Anonymization/aggregation/minimization of data

Disclosure

Did a human or machine decide?

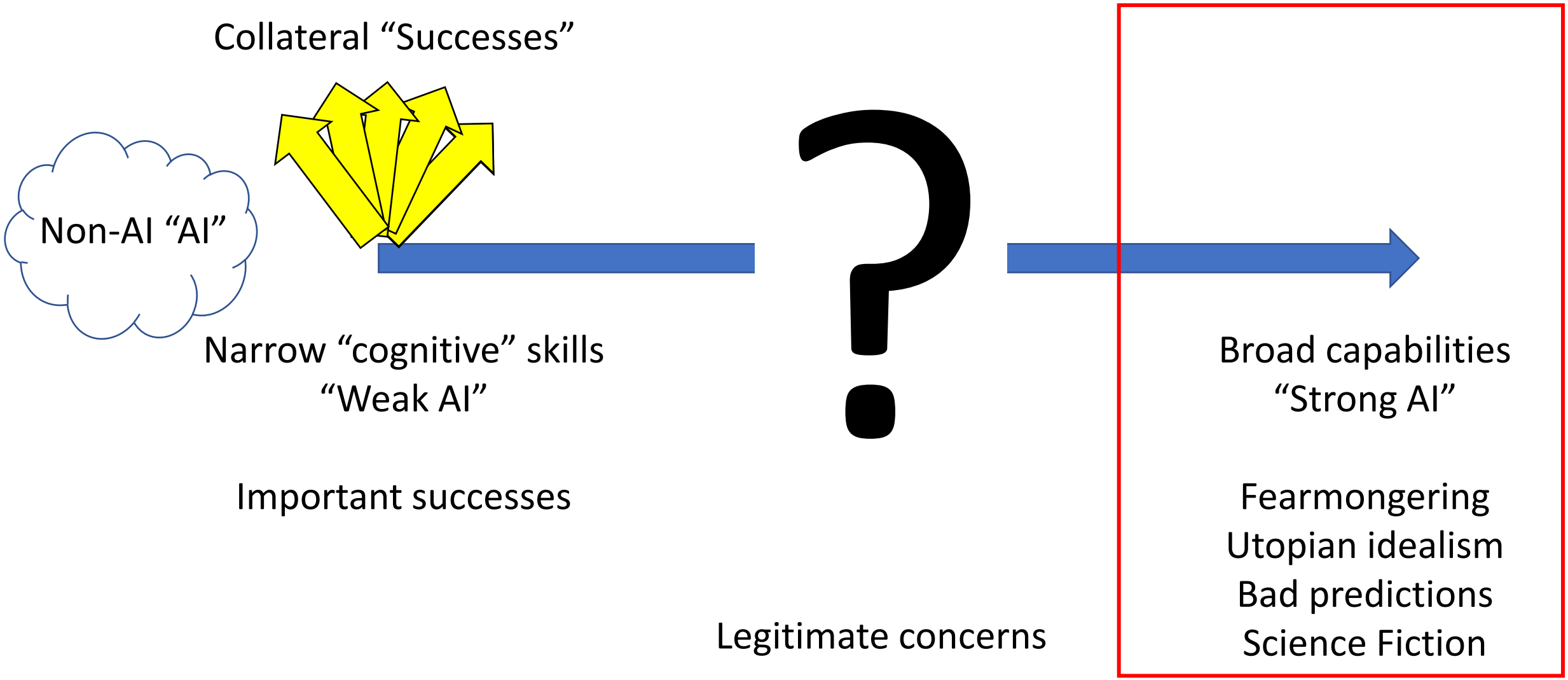
Governance/Regulation

Corporate principles

Microsoft!!!

Legal approaches

The AI Spectrum



Longer Term

We do not know how to create “artificial general intelligence” (AGI)

We are becoming more aware of the liabilities

We are thinking about how to ensure safe AGI

Singularity

Singularity

- Super-intelligent creations exceed human abilities and grow exponentially faster away from us

Singularity

- Super-intelligent creations exceed human abilities and grow exponentially faster away from us (isn’t it wonderful?)

Singularity

- Super-intelligent creations exceed human abilities and grow exponentially faster away from us (and hurt us)

Singularity

- Super-intelligent creations exceed human abilities and grow exponentially faster away from us (and hurt us)

- Three forms of speculation:
  - Machine run amok
  - Technology-enhance people run
  - Upload ourselves on machines

Singularity

- Super-intelligent creations exceed human abilities and grow exponentially faster away from us (and hurt us)

- Three forms of speculation:
  - Machine run amok
  - Technology-enhance people run
  - Upload ourselves on machines

Speculation

Is It Possible?

Philosophers of AI have debated it for decades

Is It Possible?

Philosophers of AI have debated it for decades

“Surely machines can’t do X”

The Chinese Room: It’s not really thinking

Consciousness

Embodiment

Godel’s Incompleteness Theorem

Longer Term

Generally:

We think computers can get there in some form

We don’t know how yet

We’re moving along and it’s worth working on

Longer Term

Generally:

We think computers can get there in some form

We don’t know how yet

We’re moving along and it’s worth working on

Example: AI as human collaboration vs automation

How We Talk about AI

- Differentiating automation from AGI (and where we are on each)

- How we discuss our AI work:
  - “<Such and such> could lead to <great thing>”
  - Mixing together today’s automation and AGI
  - Using human terms
    (artificial <u>intelligence</u>, reasoning, learning, …)
    (contrast with “natural language processing”)
  - Female voice for Siri, Alex, Google
  - Human name for Alexa, Watson, Siri?