Adversarial attacks that matter

Nicholas Carlini
Google

Let's attack real systems

Nicholas Carlini
Google

Why did are we studying
adversarial examples
in the first place?

Intriguing properties of neural networks

Christian Szegedy
Google Inc.

Wojciech Zaremba
New York University

Ilya Sutskever
Google Inc.

Joan Bruna
New York University

Dumitru Erhan
Google Inc.

Ian Goodfellow
University of Montreal

Rob Fergus
New York University
Facebook Inc.

5 Discussion

We demonstrated that deep neural networks have counter-intuitive properties both with respect to the semantic meaning of individual units and with respect to their discontinuities. The existence of the adversarial negatives appears to be in contradiction with the network's ability to achieve high generalization performance. Indeed, if the network can generalize well, how can it be confused by these adversarial negatives, which are indistinguishable from the regular examples? Possible explanation is that the set of adversarial negatives is of extremely low probability, and thus is never (or rarely) observed in the test set, yet it is dense (much like the rational numbers), and so it is found near every virtually every test case. However, we don't have a deep understanding of how often adversarial negatives appears, and thus this issue should be addressed in a future research.

2013

Explaining and Harnessing Adversarial Examples

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

10 Summary and Discussion

As a summary, this paper has made the following observations:

- Adversarial examples can be explained as a property of high-dimensional dot products. They are a result of models being too linear, rather than too nonlinear.
- The generalization of adversarial examples across different models can be explained as a result of adversarial perturbations being highly aligned with the weight vectors of a model, and different models learning similar functions when trained to perform the same task.
- The direction of perturbation, rather than the specific point in space, matters most. Space is not full of pockets of adversarial examples that finely tile the reals like the rational numbers.
- Because it is the direction that matters most, adversarial perturbations generalize across different clean examples.

2014

TOWARDS DEEP NEURAL NETWORK ARCHITECTURES ROBUST TO ADVERSARIAL EXAMPLES

Shixiang Gu
Panasonic Silicon Valley Laboratory
Panasonic R&D Company of America
shane.gu@us.panasonic.com

Luca Rigazio
Panasonic Silicon Valley Laboratory
Panasonic R&D Company of America
luca.rigazio@us.panasonic.com

5 CONCLUSIONS

We tested several denoising architectures to reduce the effects of the adversarial examples, and conclude that while the simple and stable structure of adversarial examples makes them easy to remove with autoencoders, the resulting stacked network is even more sensitive to new adversarial examples. We conclude that neural network's sensitivity to adversarial examples is more related to intrinsic deficiencies in the training procedure and objective function than to model topology. The crux of the problem is then to come up with an appropriate training procedure and objective function that can efficiently make the network learn flat, invariant regions around the training data. We propose Deep Contractive Networks to explicitly learn invariant features at each layer and show some positive initial results.

2014

This line of work was entirely focused on *generalization*

However another parallel direction did consider *security*

Evasion attacks against machine learning at test time

Battista Biggio[1], Igino Corona[1], Davide Maiorca[1], Blaine Nelson[2], Nedim Šrndić[3], Pavel Laskov[3], Giorgio Giacinto[1], and Fabio Roli[1]

Abstract. In security-sensitive applications, the success of machine learning depends on a thorough vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary may attempt to evade a deployed system at test time by carefully manipulating attack samples. In this work, we present a simple but effective gradient-based approach that can be exploited to systematically assess the security of several, widely-used classification algorithms against evasion attacks. Following a recently proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for the classifier by increasing the attacker's knowledge of the system and her ability to manipulate attack samples. This gives the classifier designer a better picture of the classifier performance under evasion attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our approach on the relevant security task of malware detection in PDF files, and show that such systems can be easily evaded. We also sketch some countermeasures suggested by our analysis.

2013

Evasion attacks against machine learning at test time

Battista Biggio[1], Igino Corona[1], Davide Maiorca[1], Blaine Nelson[2], Nedim Šrndić[3], Pavel Laskov[3], Giorgio Giacinto[1], and Fabio Roli[1]

Abstract. In security-sensitive applications, the success of machine learning depends on a thorough vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary may attempt to evade a deployed system at test time by carefully manipulating attack samples. In this work, we present a simple but effective gradient-based approach that can be exploited to systematically assess the security of several, widely-used classification algorithms against evasion attacks. Following a recently proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for the classifier by increasing the attacker's knowledge of the system and her ability to manipulate attack samples. This gives the classifier designer a better picture of the classifier performance under evasion attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our approach on the relevant security task of malware detection in PDF files, and show that such systems can be easily evaded. We also sketch some countermeasures suggested by our analysis.

2013

This talk:

Do we have real attacks yet?

POLICY FORUM

MACHINE LEARNING

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

By **Samuel G. Finlayson[1], John D. Bowers[2], Joichi Ito[3], Jonathan L. Zittrain[2], Andrew L. Beam[4], Isaac S. Kohane[1]**

Understanding Adversarial Attacks on Deep Learning Based
Medical Image Analysis Systems

c Lin Gu [d] Yisen Wang [e] Yitian Zhao [f] James Bailey [b] Feng Lu [**, a, c]
Technology and Systems, School of CSE, Beihang University, Beijing, China.
rmation Systems, The University of Melbourne, Parkville, VIC 3010, Australia.
enter for Big Data-Based Precision Medicine, Beihang University, Beijing, China.
National Institute of Informatics, Tokyo 101-8430, Japan.
[e]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.
[f]Cixi Instuitue of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo, China.

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, *IEEE* and M. Shamim Hossain, Senior Member, *IEEE,* Nabil A. Alrajeh, Fawaz Alsolami

Advers
learning

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou[1 [0000−0001−5127−3170]], Konstantinos Sechidis[1 [0000−0001−6582−7453]], James Weatherall[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.papangelou,konstantinos.sechidis, gavin.brown}@manchester.ac.uk
[2] Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK
james.weatherall@astrazeneca.com

and Robust Machine Learning for Healthcare: A Survey

Qayyum[1], Junaid Qadir[1], Muhammad Bilal[2], and Ala Al-Fuqaha[3*]
ormation Technology University (ITU), Punjab, Lahore, Pakistan
niversity of the West England (UWE), Bristol, United Kingdom
[3] Hamad Bin Khalifa University (HBKU), Doha, Qatar

- ☐ New ideas
- ☐ Real system
- ☐ Threat model

- [x] New ideas
- [ ] Real system
- [ ] Threat model

- [x] New ideas
- [ ] Real system
- [ ] Threat model

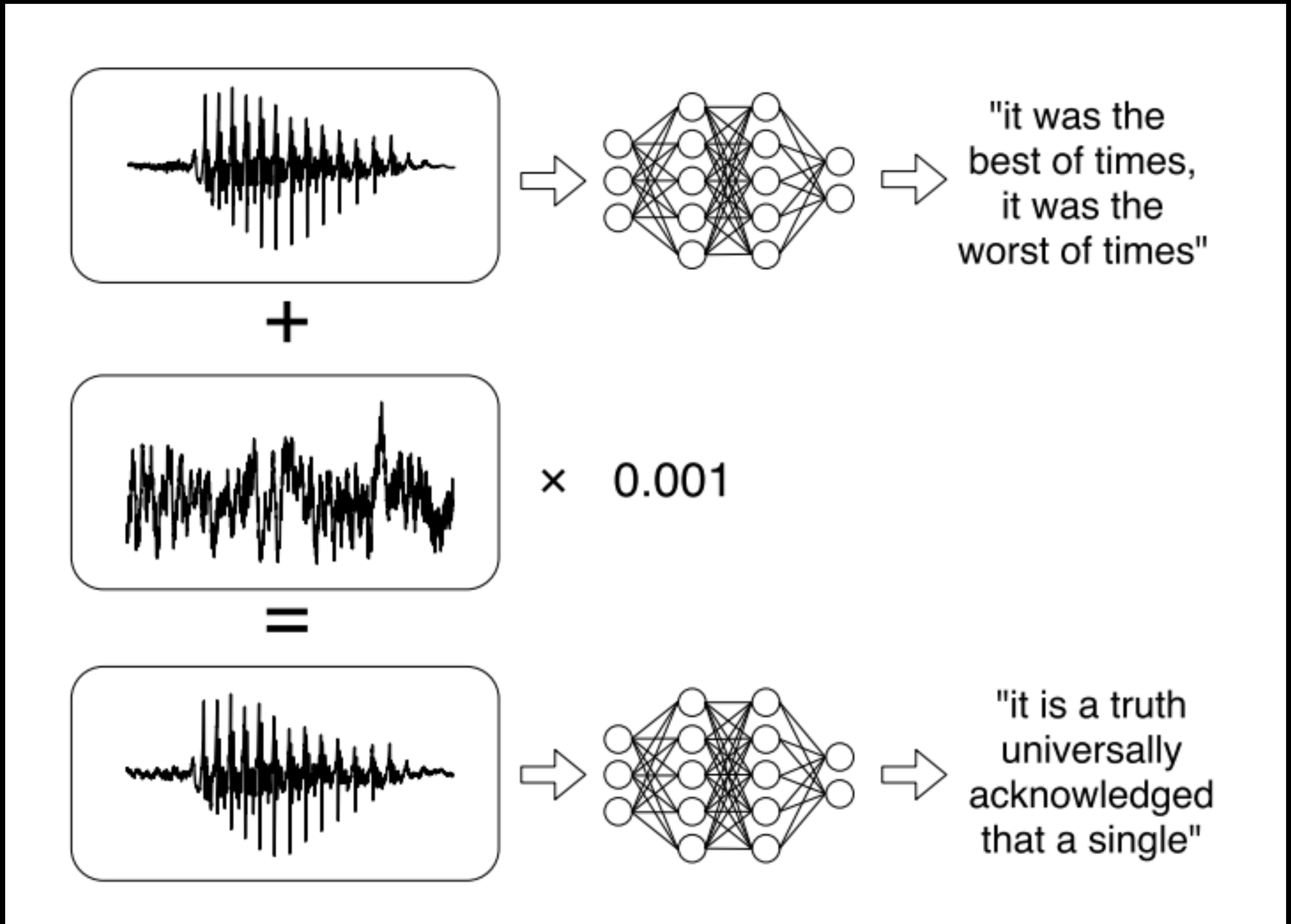
- ☐ New ideas
- ☐ Real system
- ☐ Threat model

- ☐ New ideas
- ☐ Real system
- ☐ Threat model

- ✓ New ideas
- ✗ Real system
- ✗ Threat model

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

Nicholas Carlini David Wagner
University of California, Berkeley



- [x] New ideas
- [ ] Real system
- [ ] Threat model

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]

- [x] New ideas
- [ ] Real system
- [ ] Threat model

Experimental Security Research of Tesla Autopilot

Tencent Keen Security Lab

2019-03

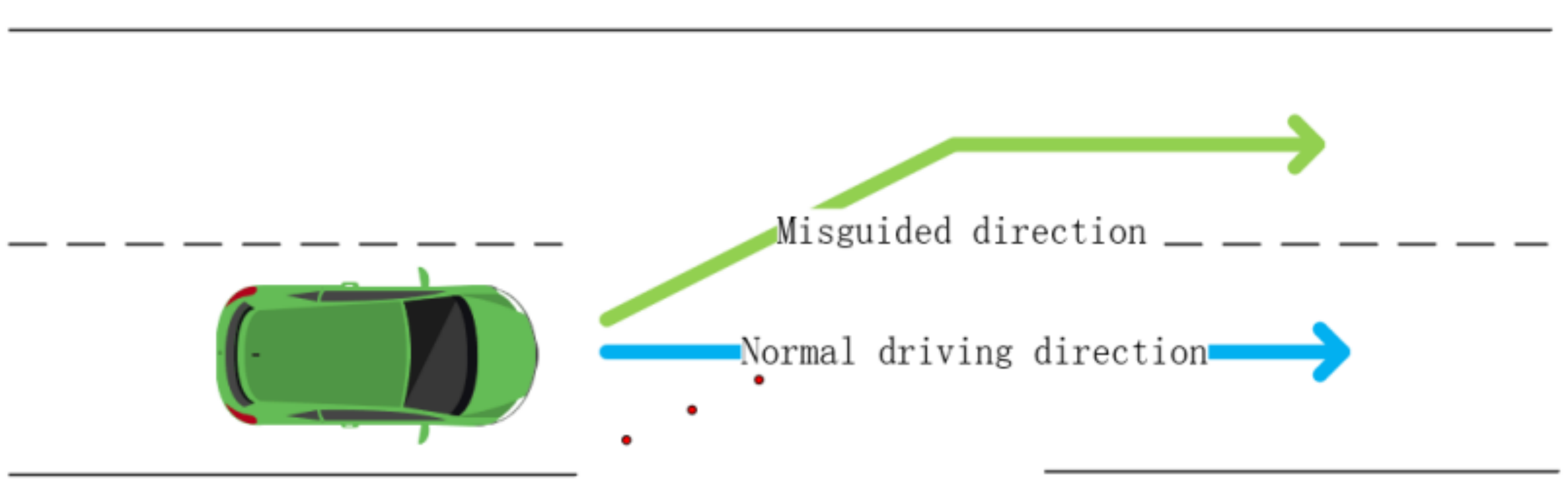


Fig 34. Fake lane mode in physical world

Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations

Pengfei Jing[1,2], Qiyi Tang[2], Yuefeng Du[2], Lei Xue[1], Xiapu Luo[1]*, Ting Wang[3], Sen Nie[2], Shi Wu[2]

[1]Department of Computing, The Hong Kong Polytechnic University
[2]Keen Security Lab, Tencent
[3]College of Information Sciences and Technology, Pennsylvania State University

utopilot

Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack

Takami Sato*
UC Irvine
takamis@uci.edu

Junjie Shen*
UC Irvine
junjies1@uci.edu

Ningfei Wang
UC Irvine
ningfei.wang@uci.edu

Yunhan Jia
ByteDance
yunhan.jia@bytedance.com

Xue Lin
Northeastern University
xue.lin@northeastern.edu

Qi Alfred Chen
UC Irvine
alfchen@uci.edu

Fig 34. Fake lane mode in physical world

- [x] New ideas
- [x] Real system
- [ ] Threat model

Motivating the Rules of the Game for Adversarial Example Research

Justin Gilmer[1]*, Ryan P. Adams[2], Ian Goodfellow[1],
David Andersen[1], George E. Dahl[1]*
{gilmer, goodfellow, dga, gdahl}@google.com, rpa@princeton.edu
[1]Google Brain; [2]Princeton

July 2018

Abstract

Advances in machine learning have led to broad deployment of systems with impressive performance on important problems. Nonetheless, these systems can be induced to make errors on data that are surprisingly similar to examples the learned system handles correctly. The existence of these errors raises a variety of questions about out-of-sample generalization and whether bad actors might use such examples to abuse deployed systems. As a result of these security concerns, there has been a flurry of recent papers proposing algorithms to defend against such malicious perturbations of correctly handled examples. It is unclear how such misclassifications represent a different kind of security problem than other errors, or even other attacker-produced examples that have no specific relationship to an uncorrupted input. In this paper, we argue that adversarial example defense papers have, to date, mostly considered abstract, toy games that do not relate to any specific security concern. Furthermore, defense papers have not yet precisely described all the abilities and limitations of attackers that would be relevant in practical security. Towards this end, we establish a taxonomy of motivations, constraints, and abilities for more plausible adversaries. Finally, we provide a series of recommendations outlining a path forward for future work to more clearly articulate the threat model and perform more meaningful evaluation.

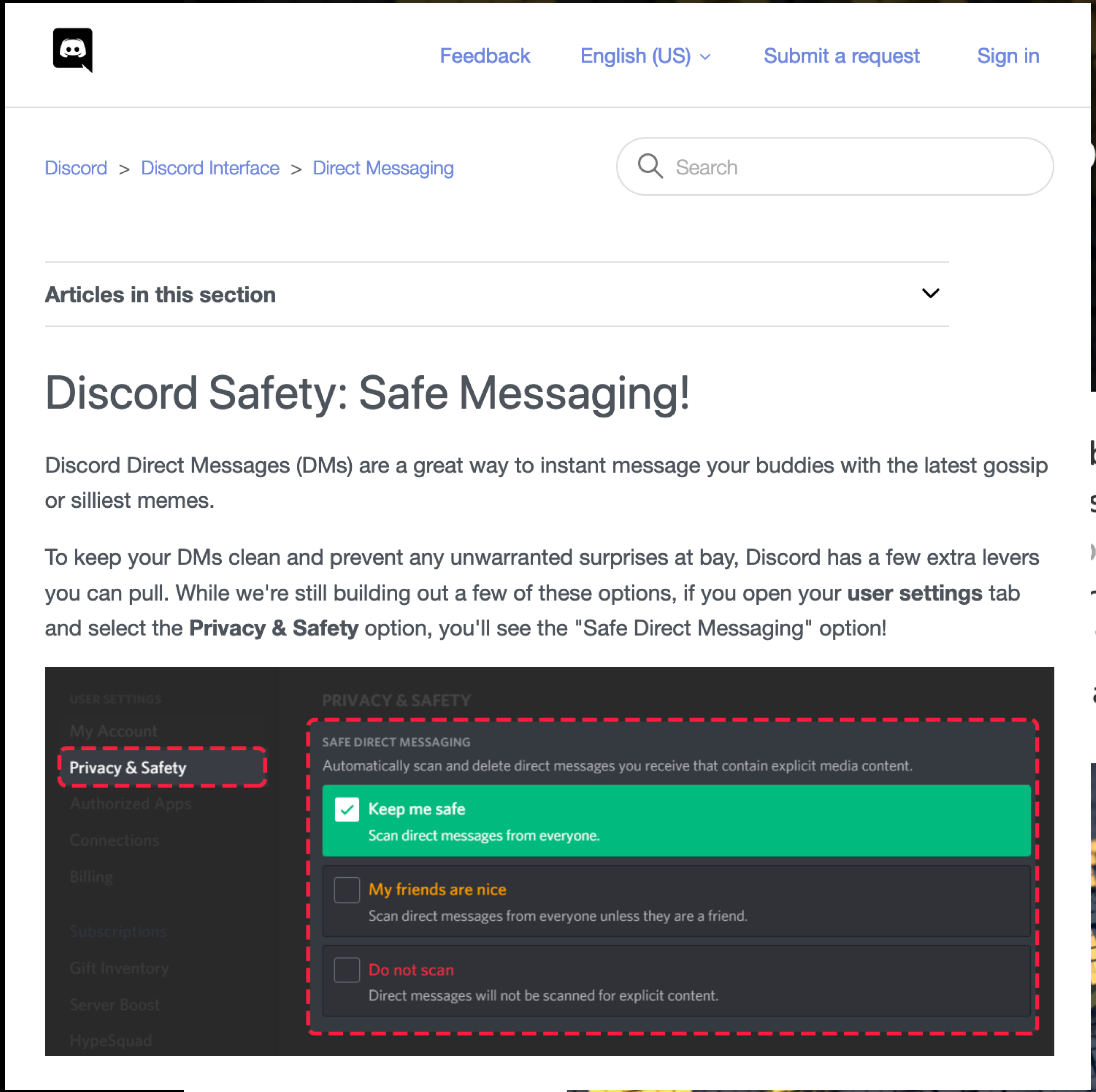
We're now really good at generating adversarial examples.

What's next?

Let's attack real systems

(that have realistic threat models)

Content Filtering



Feedback English (US) Submit a request Sign in

Discord > Discord Interface > Direct Messaging

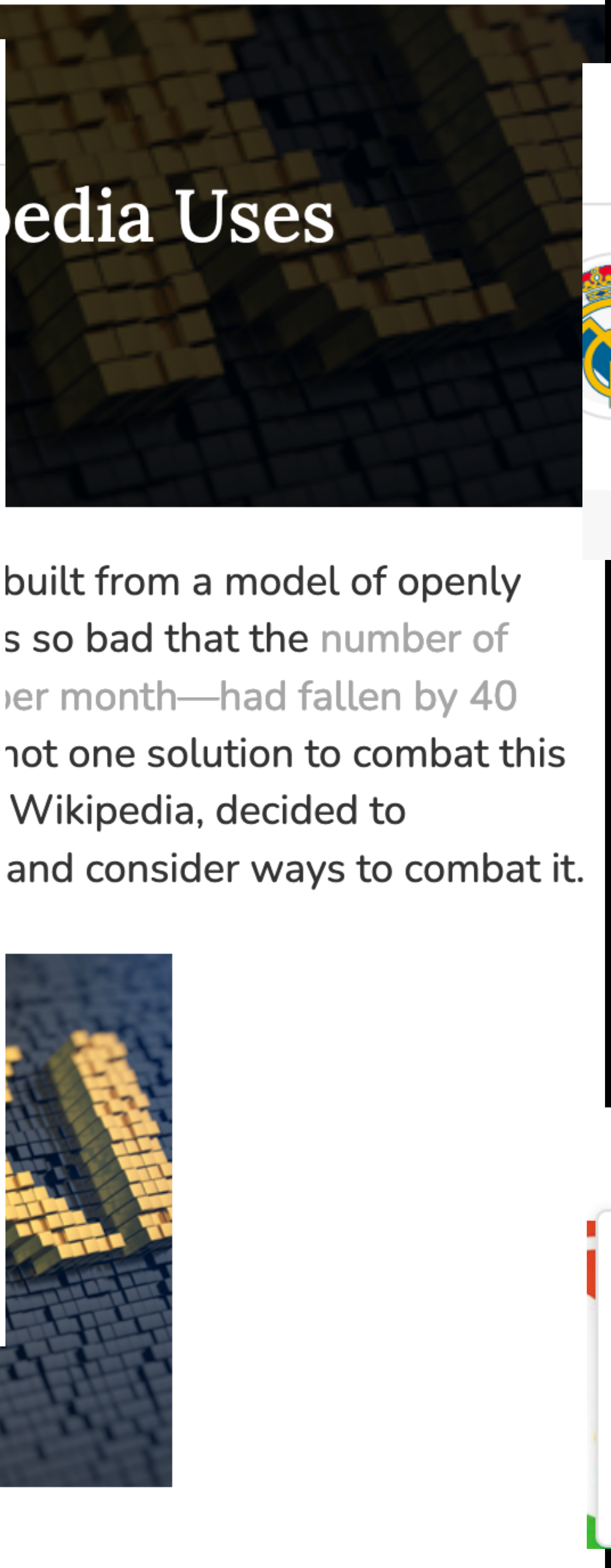
Search

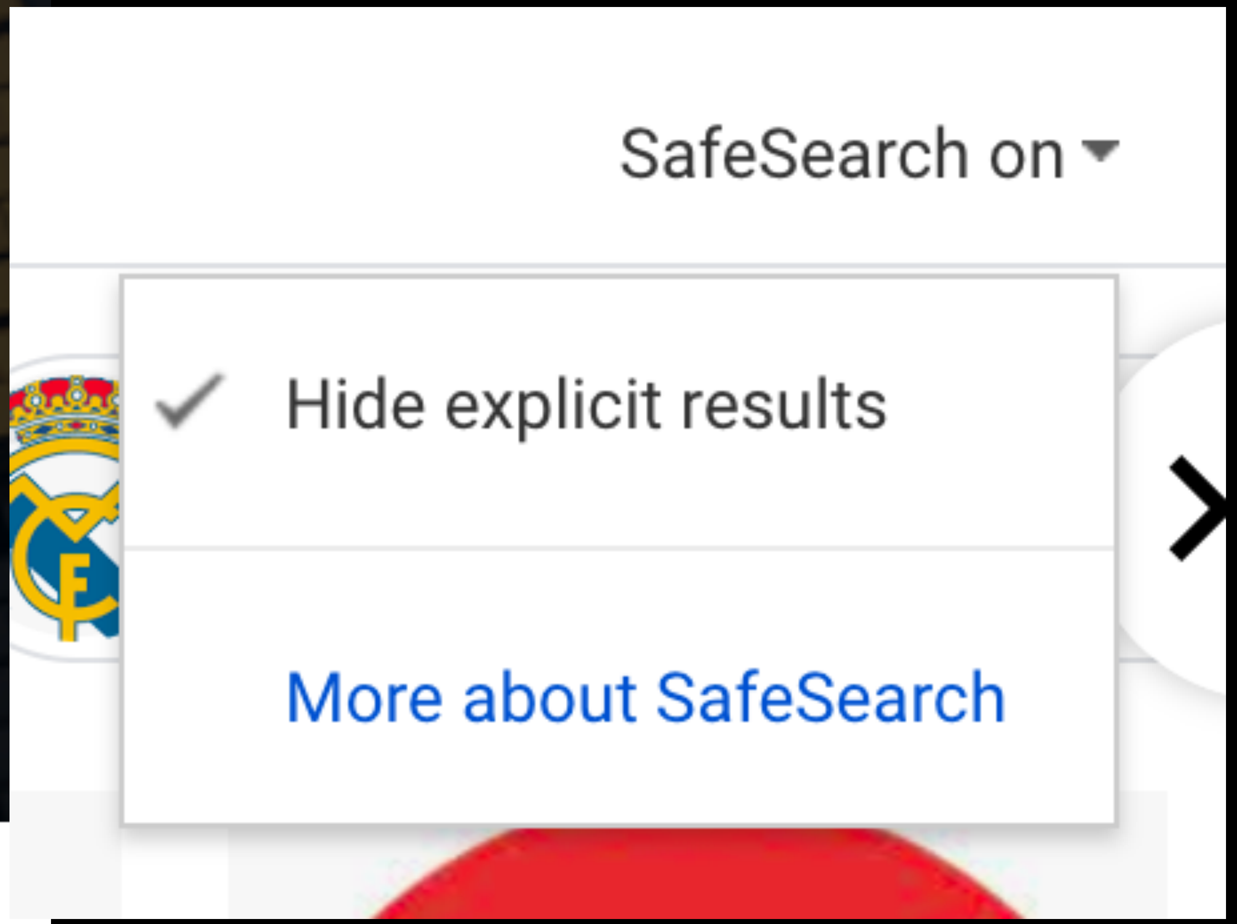
Articles in this section

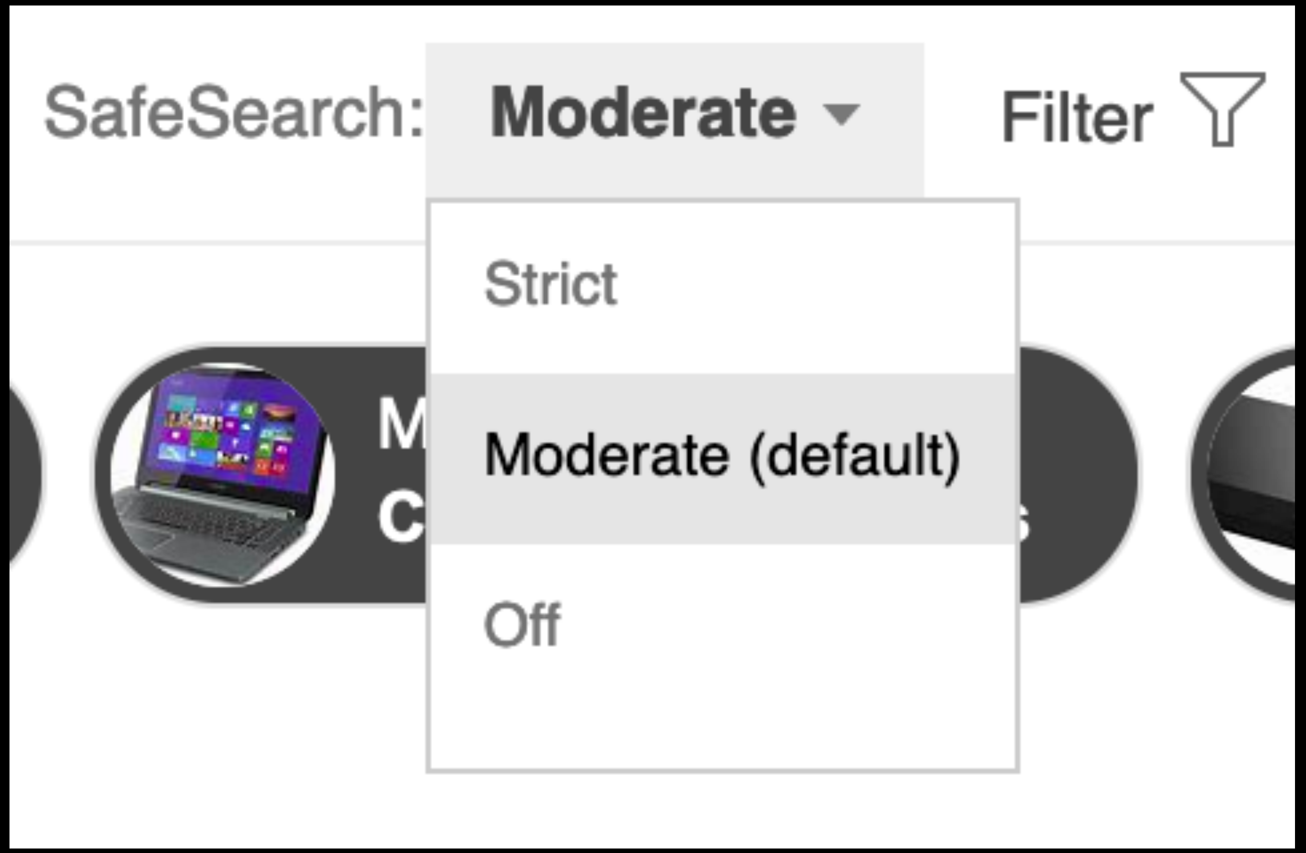
Discord Safety: Safe Messaging!

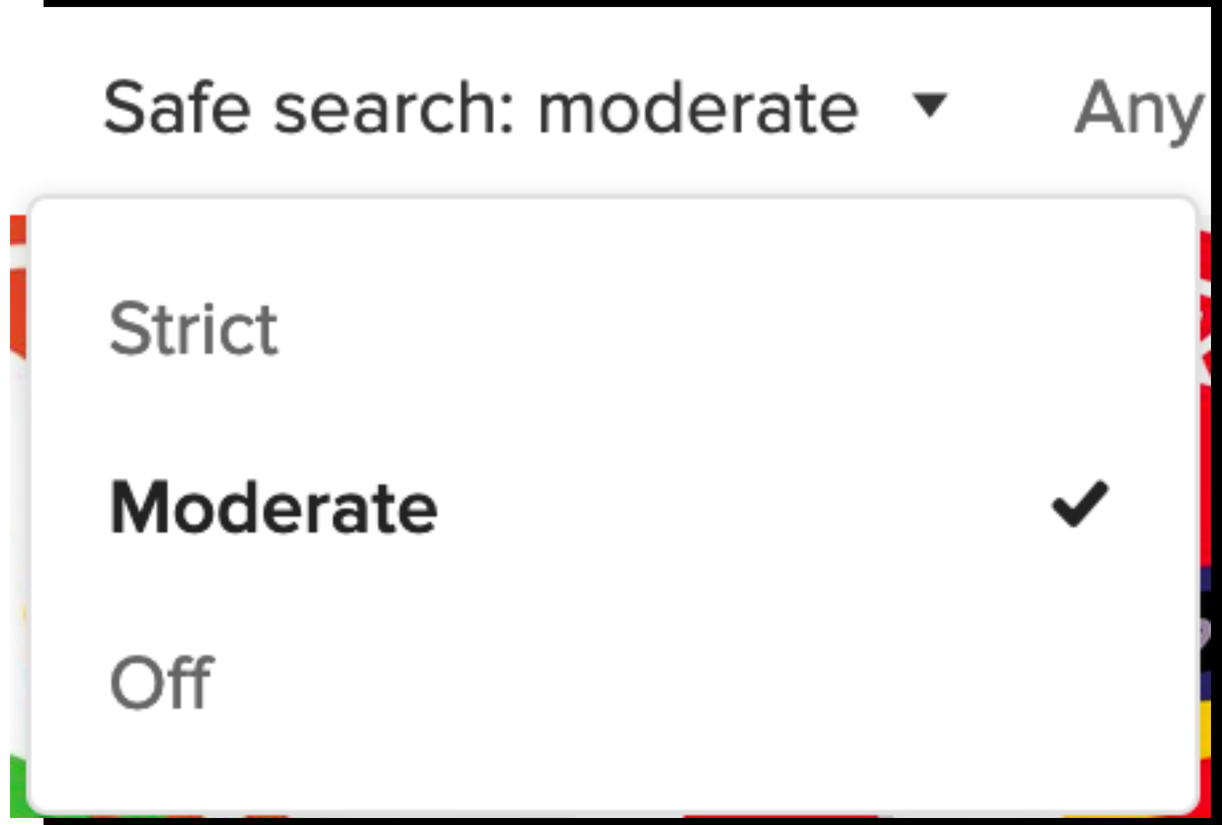
Discord Direct Messages (DMs) are a great way to instant message your buddies with the latest gossip or silliest memes.

To keep your DMs clean and prevent any unwarranted surprises at bay, Discord has a few extra levers you can pull. While we're still building out a few of these options, if you open your **user settings** tab and select the **Privacy & Safety** option, you'll see the "Safe Direct Messaging" option!









BEHAVIOR | OPINION

How AI Is Learning to Identify Toxic Online Content

Machine-learning systems could help flag hateful, threatening or offensive language

By Laura Hanu, James Thewlis, Sasha Haco on February 8, 2021 أعرض هذا باللغة العربية

Facebook

Update on Our Progress on AI and Hate Speech Detection

February 11, 2021
By Mike Schroepfer, Chief Technology Officer

ORES

ORES is a web service that provides machine learning as a
like Wikipedia and Wikidata. The system is designed to hel
wiki-work and to increase their productivity by automating
and removing edits made in bad faith. ORES is developed
team that specializes in building transparent, auditable, op
intelligence (AI) to support human decision-making.

ORES is intended to be used as a source of structured info
developers and product developers at the Wikimedia Foundation and Wikimedia Deutschland. Most users access ORES via 3rd party tools like Huggle and Special:RecentChanges on Wikimedia wikis. To access ORES scores, a simple scores API and a reference UI are available.

Tune. —Experimental

Control the comments you see on YouTube, Twitter, Facebook, Reddit, and Disqus.

Using machine learning to reduce toxicity online

Perspective API can help mitigate toxicity and ensure healthy dialogue online.

HOW IT WORKS →

What's potentially new:

- Limited query-only access to classifier
- Unknown network architecture
- Unknown image processing pipeline
- ???????

Malware

Malware detection through artificial intelligence and neural networks?

12. November, 2019

World First
Visual AI Based Malware Detection

The first solution that converts files into graphical representations and check whether malware is contained or not. We provide user-friendly, efficient and secure malware detection technology.
You don't want to read any further but want to test it directly? Visit our free community version at Malware.AI

Test Now | More Information

BY YIHUA LIAO | SEP 02 2021

AI/ML for Malware Detection

This is the fourth in an ongoing series of blogs focused on AI/ML.

Machine Learning for Malware Detection

Sophos AI

Pushing the boundaries of machine learning for information security

kaspersky

Learn more on kaspersky.com
#bringonthefuture

What's potentially new:

- Almost no query-only access
- Unknown feature extraction
- Unknown machine learning model
- Lp perturbations don't matter
- ???????

Ad blocking

AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning

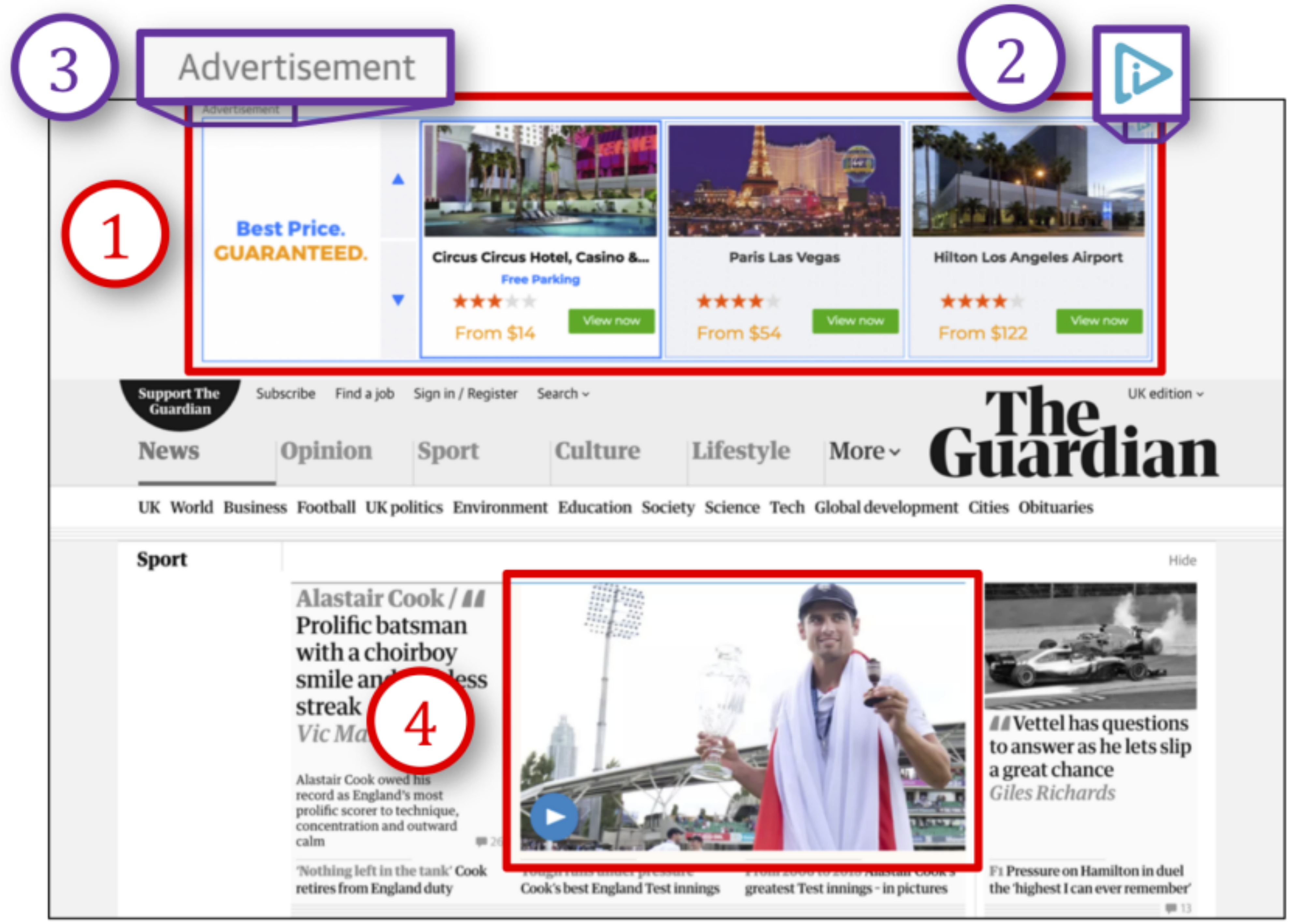
Florian Tramèr
tramer@cs.stanford.edu
Stanford University

Pascal Dupré
s9padupr@stud.uni-saarland.de
CISPA Helmholtz Center for Information Security

Gili Rusak
gili@stanford.edu
Stanford University

Giancarlo Pellegrino
gpellegrino@cispa.saarland
Stanford University, CISPA Helmholtz Center for Information Security

Dan Boneh
dabo@cs.stanford.edu
Stanford University



- [x] New ideas
- [x] Real system
- [x] Threat model

I'm sure I'm missing a lot!

defend

Let's ~~attack~~ real systems

Can we make new assumptions that are true in practice but haven't been studied extensively?

I don't care if a defense
is robust.

I care that we
learn something new


Feedback
English (US)
Submit a request
Sign in
Discord > Discord Interface > Direct Messaging
Search
Articles in this section
Discord Safety: Safe Messaging!
Discord Direct Messages (DMs) are a great way to instant message your buddies with the latest gossip or silliest memes.
To keep your DMs clean and prevent any unwarranted surprises at bay, Discord has a few extra levers you can pull. While we're still building out a few of these options, if you open your user settings tab and select the Privacy & Safety option, you'll see the "Safe Direct Messaging" option!
PRIVACY & SAFETY
SAFE DIRECT MESSAGING
Automatically scan and delete direct messages you receive that contain explicit media content.
Privacy & Safety
Keep me safe
Scan direct messages from everyone.
My friends are nice
Scan direct messages from everyone unless they are a friend.
Do not scan
Direct messages will not be scanned for explicit content.
edia Uses
built from a model of openly
s so bad that the number of
er month—had fallen by 40
not one solution to combat this
Wikipedia, decided to
and consider ways to combat it.
SafeSearch on
Hide explicit results
More about SafeSearch
SafeSearch: Moderate
Filter
Strict
Moderate (default)
Off
Safe search: moderate
Any
Strict
Moderate
Off

BEHAVIOR | OPINION

How AI Is Learning to Identify Toxic Online Content

Machine-learning systems could help flag hateful, threatening or offensive language

By Laura Hanu, James Thewlis, Sasha Haco on February 8, 2021 أعرض هذا باللغة العربية

Facebook

Update on Our Progress on AI and Hate Speech Detection

February 11, 2021
By Mike Schroepfer, Chief Technology Officer

ORES

ORES is a web service that provides machine learning as a
like Wikipedia and Wikidata. The system is designed to hel
wiki-work and to increase their productivity by automating
and removing edits made in bad faith. ORES is developed
team that specializes in building transparent, auditable, op
intelligence (AI) to support human decision-making.

ORES is intended to be used as a source of structured info
developers and product developers at the Wikimedia Foundation and Wikimedia Deutschland. Most users access ORES via 3rd party tools like Huggle and Special:RecentChanges on Wikimedia wikis. To access ORES scores, a simple scores API and a reference UI are available.

Tune. —Experimental

Control the comments you see on YouTube, Twitter, Facebook, Reddit, and Disqus.

Using machine learning to reduce toxicity online

Perspective API can help mitigate toxicity and ensure healthy dialogue online.

HOW IT WORKS →

Malware detection through artificial intelligence and neural networks?

12. November, 2019

World First
Visual AI Based Malware Detection

The first solution that converts files into graphical representations and check whether malware is contained or not. We provide user-friendly, efficient and secure malware detection technology.
You don't want to read any further but want to test it directly? Visit our free community version at Malware.AI

Test Now | More Information

BY YIHUA LIAO | SEP 02 2021

AI/ML for Malware Detection

This is the fourth in an ongoing series of blogs focused on AI/ML.

Machine Learning for Malware Detection

Sophos AI

Pushing the boundaries of machine learning for information security

kaspersky

Learn more on kaspersky.com
#bringonthefuture

AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning

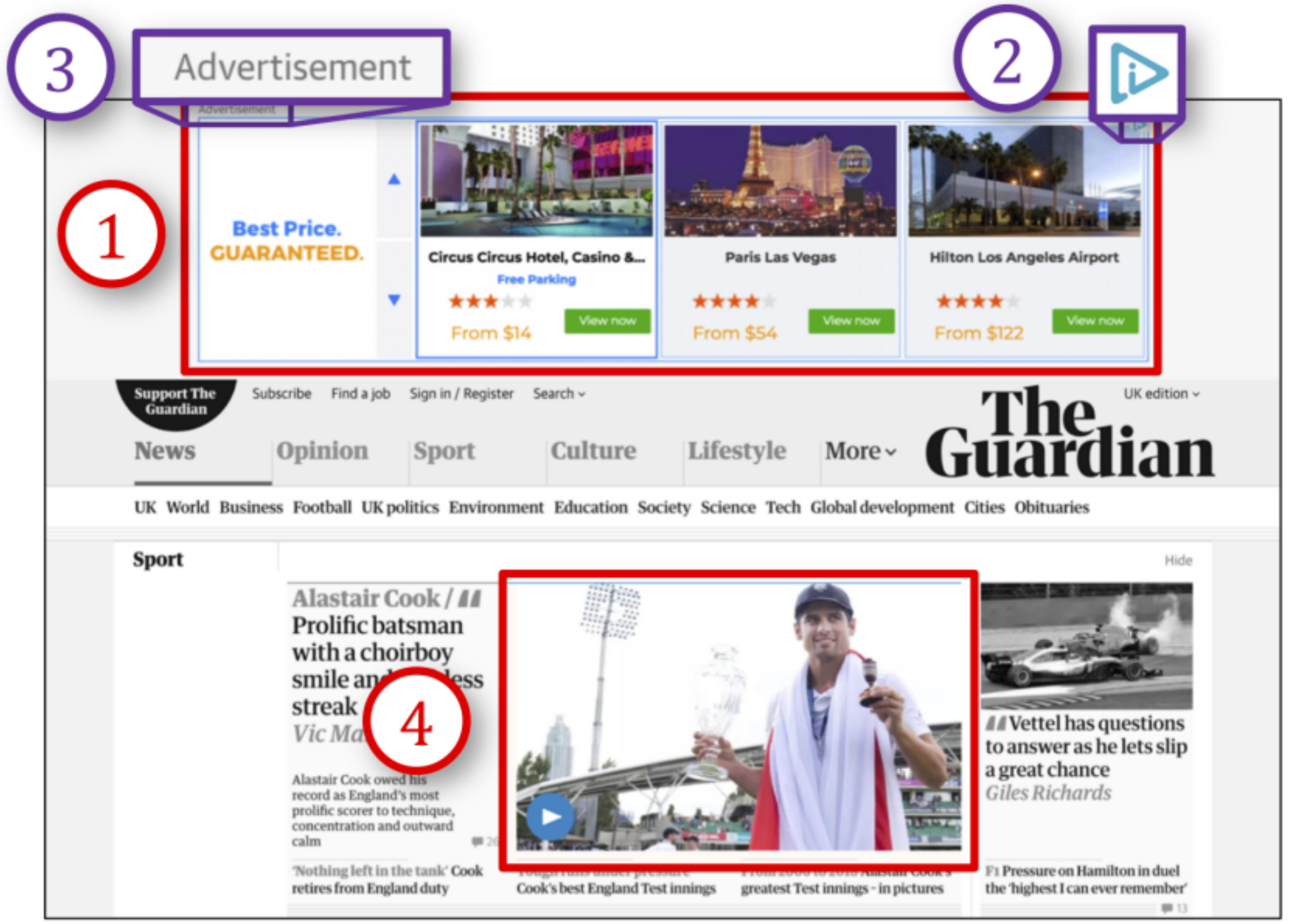
Florian Tramèr
tramer@cs.stanford.edu
Stanford University

Pascal Dupré
s9padupr@stud.uni-saarland.de
CISPA Helmholtz Center for Information Security

Gili Rusak
gili@stanford.edu
Stanford University

Giancarlo Pellegrino
gpellegrino@cispa.saarland
Stanford University, CISPA Helmholtz Center for Information Security

Dan Boneh
dabo@cs.stanford.edu
Stanford University



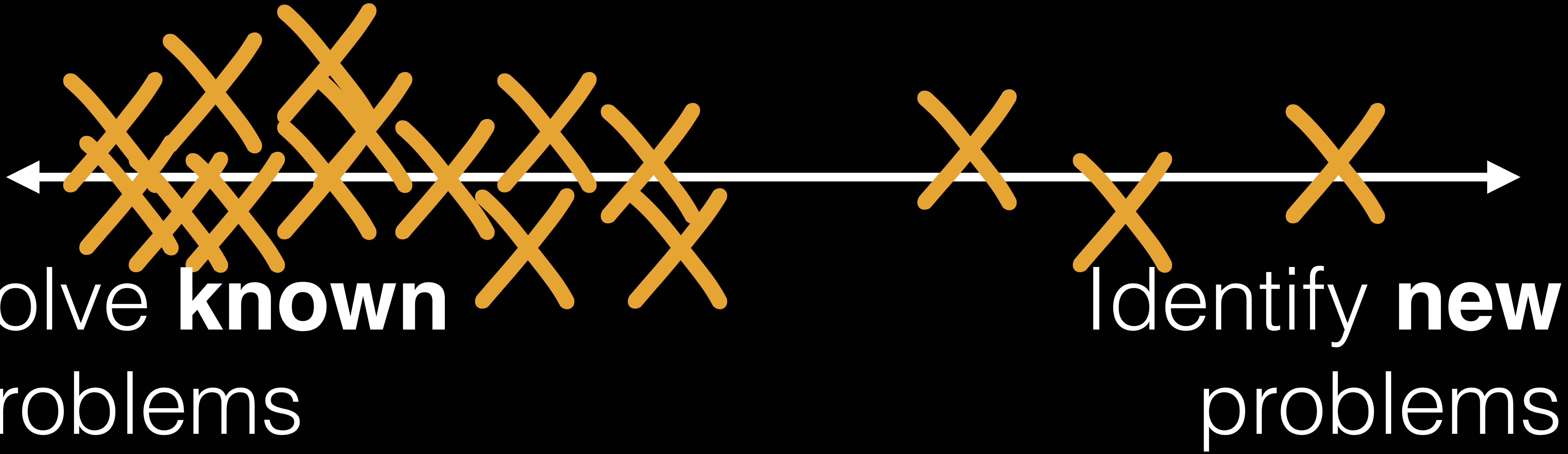
Conclusion

Two types of research

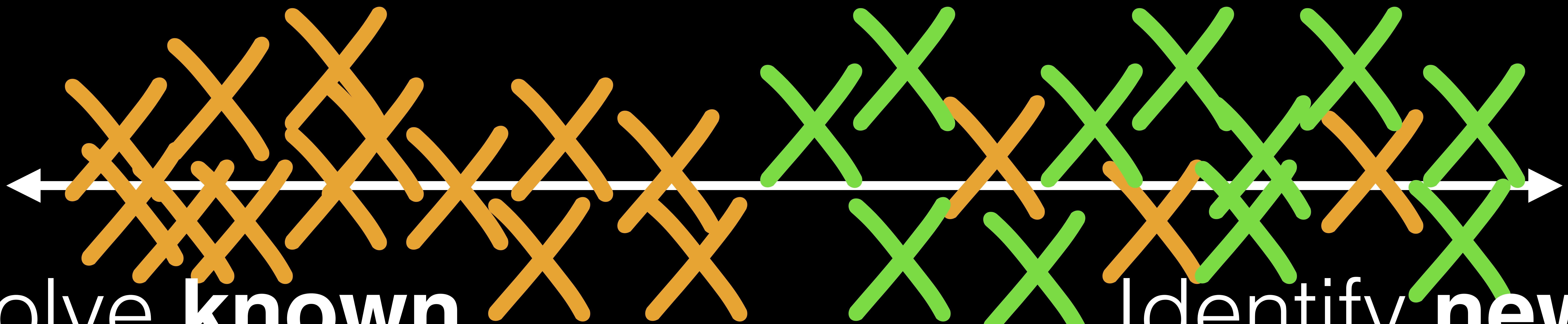
Solve **known** problems

Identify **new** problems

In adversarial machine learning:



In adversarial machine learning:



Solve **known** problems

Identify **new** problems

By studying real systems,
we can better discover the
limitations of our current tools