Effective Altruism
Foundation
http://ea-foundation.org

An Overview of the AI Safety Landscape

Workshop on Reliable Artificial Intelligence 2017, ETH Zurich

Max Daniel

Research Project Manager, Effective Altruism Foundation

https://blog.openai.com/faulty-reward-functions/

https://blog.openai.com/faulty-reward-functions/

Concrete Problems in AI Safety

Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

"[C]oncrete safety problems that are ready for experimentation today and relevant to the cutting edge of AI systems"

1. Avoid negative side effects
2. Avoid reward hacking
3. Scalable oversight
4. Safe exploration
5. Robustness to distributional shift

Amodei, Olah et al. 2016

Algorithms for Inverse Reinforcement Learning

Andrew Y. Ng ANG@CS.BERKELEY.EDU
Stuart Russell RUSSELL@CS.BERKELEY.EDU
Computer Science Division, U.C. Berkeley, Berkeley, CA 94720 USA

Cooperative Inverse Reinforcement Learning

Dylan Hadfield-Menell* **Anca Dragan** **Pieter Abbeel** **Stuart Russell**
Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94709

Ng and Russell (ICML 2000), Hadfield-Menell et al. (NIPS 2016)

Deep Reinforcement Learning from Human Preferences

Paul F Christiano
OpenAI
paul@openai.com

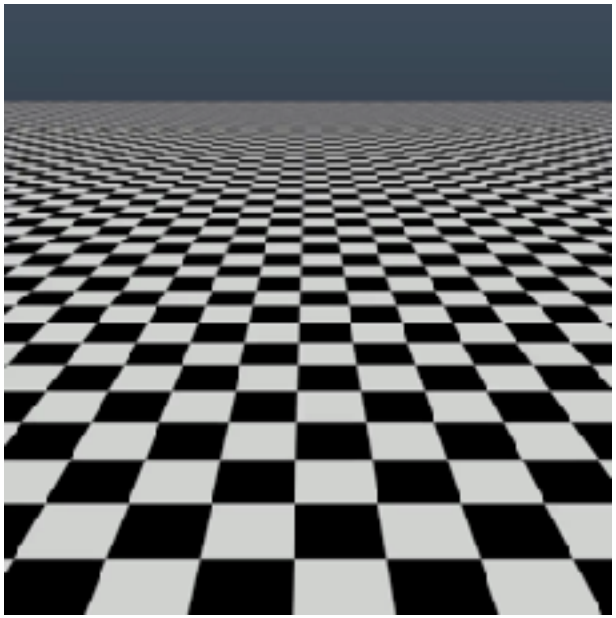
Jan Leike
DeepMind
leike@google.com

Tom B Brown
nottombrown@gmail.com

Miljan Martic
DeepMind
miljanm@google.com

Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com



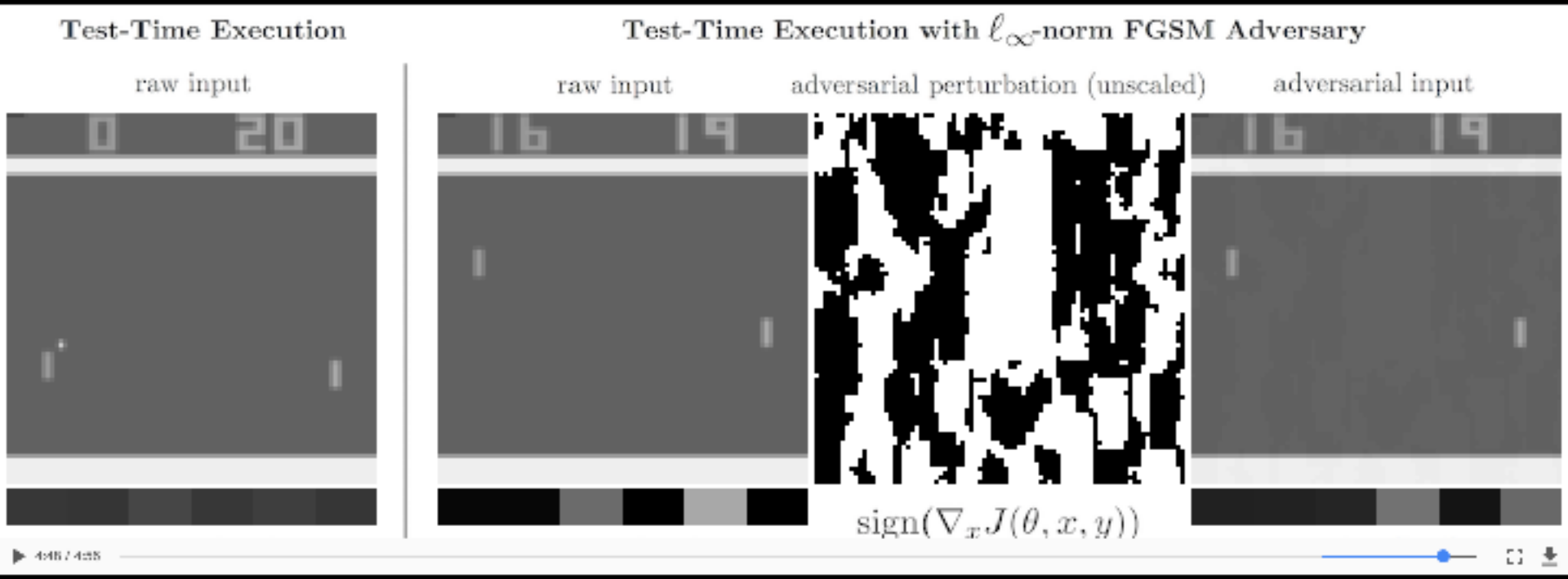
Christiano et al. 2017

Security

Adversarial Attacks on Neural Network Policies

Sandy Huang[†], **Nicolas Papernot**[‡], **Ian Goodfellow**[§], **Yan Duan**[†§], **Pieter Abbeel**[†§]
[†] University of California, Berkeley, Department of Electrical Engineering and Computer Sciences
[‡] Pennsylvania State University, School of Electrical Engineering and Computer Science
[§] OpenAI

Huang et al. 2017



Source: http://rll.berkeley.edu/adversarial/videos/pong_a3c_trpo_l-inf.mp4

Corrigibility

Corrigibility

In *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015.* AAAI Publications.

Nate Soares and **Benja Fallenstein** and **Eliezer Yudkowsky**
Machine Intelligence Research Institute
{nate,benja,eliezer}@intelligence.org

Stuart Armstrong
Future of Humanity Institute
University of Oxford
stuart.armstrong@philosophy.ox.ac.uk

Safely Interruptible Agents*

Laurent Orseau
Google DeepMind
5 New Street Square,
London EC4A 3TW, UK
lorseau@google.com

Stuart Armstrong
The Future of Humanity Institute
University of Oxford, UK
stuart.armstrong@philosophy.ox.ac.uk
Machine Intelligence Research Institute
Berkeley, CA 94704

Soares et al. (AAAI 2015), Orseau and Armstrong (UAI 2016)

Privacy

SEMI-SUPERVISED KNOWLEDGE TRANSFER FOR DEEP LEARNING FROM PRIVATE TRAINING DATA

Nicolas Papernot*
Pennsylvania State University
ngp5056@cse.psu.edu

Martín Abadi
Google Brain
abadi@google.com

Úlfar Erlingsson
Google
ulfar@google.com

Ian Goodfellow
Google Brain†
goodfellow@google.com

Kunal Talwar
Google Brain
kunal@google.com

Papernot et al. (ICLR 2017)

Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda

In *The Technological Singularity: Managing the Journey*. Springer. 2017

Nate Soares and **Benya Fallenstein**
Machine Intelligence Research Institute
{nate,benya}@intelligence.org

"This technical agenda primarily covers topics that the authors believe are *tractable, uncrowded, focused,* and *unable to be outsourced* to forerunners of the target AI system."

1. Realistic World-Models
2. Decision Theory
3. Logical Uncertainty
4. Vingean Reflection

1) Research Goal
2) Research Funding
3) Science-Policy Link
4) Research Culture
5) Race Avoidance
6) Safety
7) Failure Transparency
8) Judicial Transparency
9) Responsibility
10) Value Alignment
11) Human Values
12) Personal Privacy
13) Liberty and Privacy
14) Shared Benefit
15) Shared Prosperity
16) Human Control
17) Non-subversion
18) AI Arms Race
19) Capability Caution
20) Importance
21) Risks
22) Recursive Self-Improvement
23) Common Good

Source: Asilomar AI Principles

Conclusion

- Ensuring that AI agents do what we want is a nontrivial problem.
- Technical AI safety is a thriving field in AI/ML research.
- Several research agendas and concrete problems have been pursued.
- Complements contributions from law, economics, policy, philosophy, social science, …

Thank you.

max.daniel@ea-foundation.org