Future of Humanity Institute
UNIVERSITY OF OXFORD

STRATEGIC ARTIFICIAL INTELLIGENCE RESEARCH CENTRE

UNIVERSITY OF OXFORD

INTERACTIONS BETWEEN THE AI CONTROL PROBLEM AND THE GOVERNANCE PROBLEM

Professor Nick Bostrom

University of Oxford
Director, Strategic Artificial Intelligence Research Center
Director, Future of Humanity Institute

Q:

“If human level general AI is developed, then what are likely outcomes?”

Google DeepMind Challenge Match 4국
LIVE
바둑TV
00:59:42
이세돌 0 VS 3 알파고
ALPHAGO
01:32:15
SAM IK
SFC MALL

Lee Sedol v. AlphaGO

Oct 2015 “Based on its level seen in the match (against Fan), I think I will win the game by a near landslide”

Feb 2016 “I have heard that Google DeepMind’s AI is surprisingly strong and getting stronger, but I am confident that I can win at least this time”

Mar 9, 2016 “I was very surprised because I didn’t think I would lose”

Mar 10, 2016 “I’m quite speechless ... I am in shock. I can admit that ... the third game is not going to be easy for me”

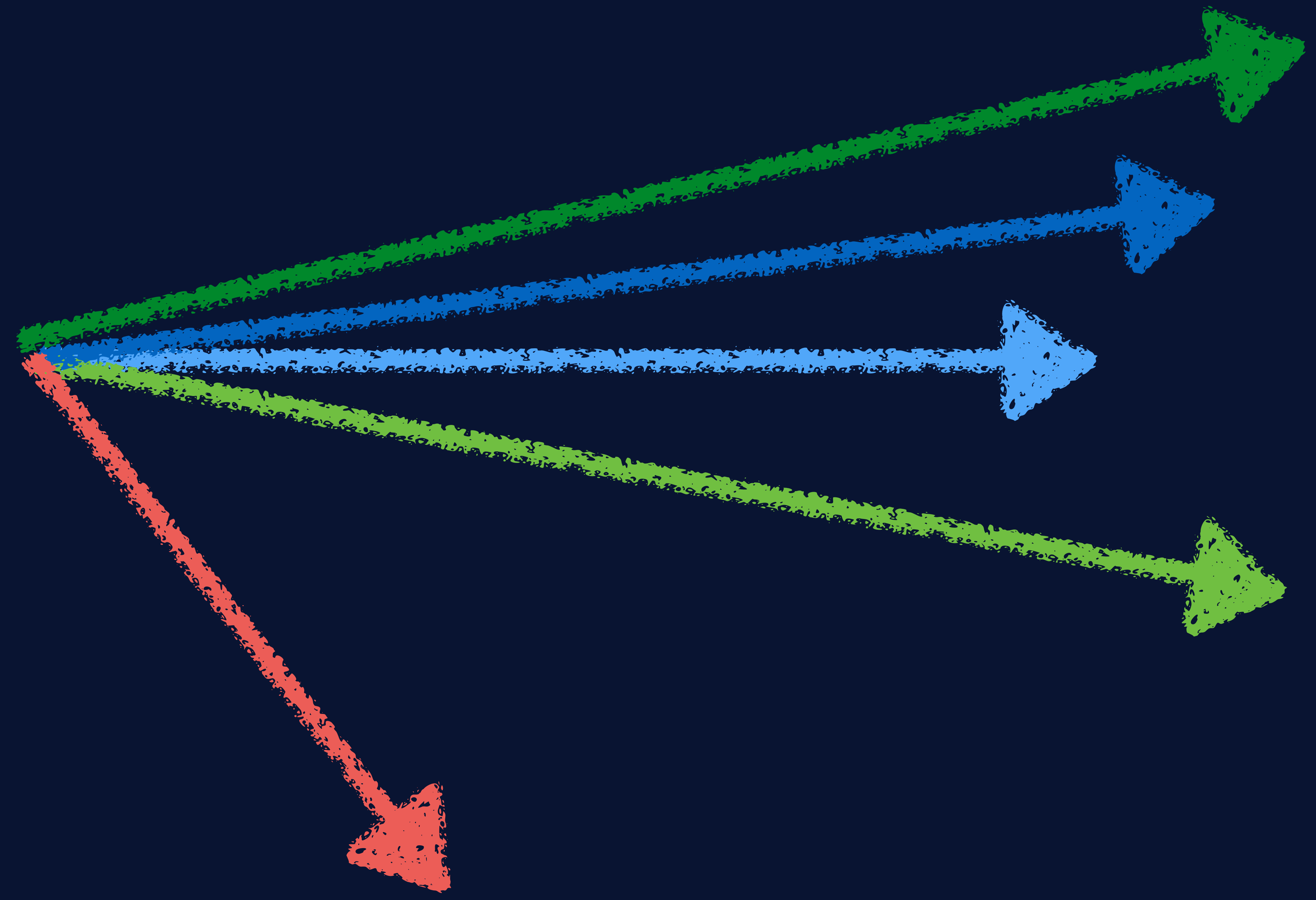
Mar 12, 2016 “I kind of felt powerless.”

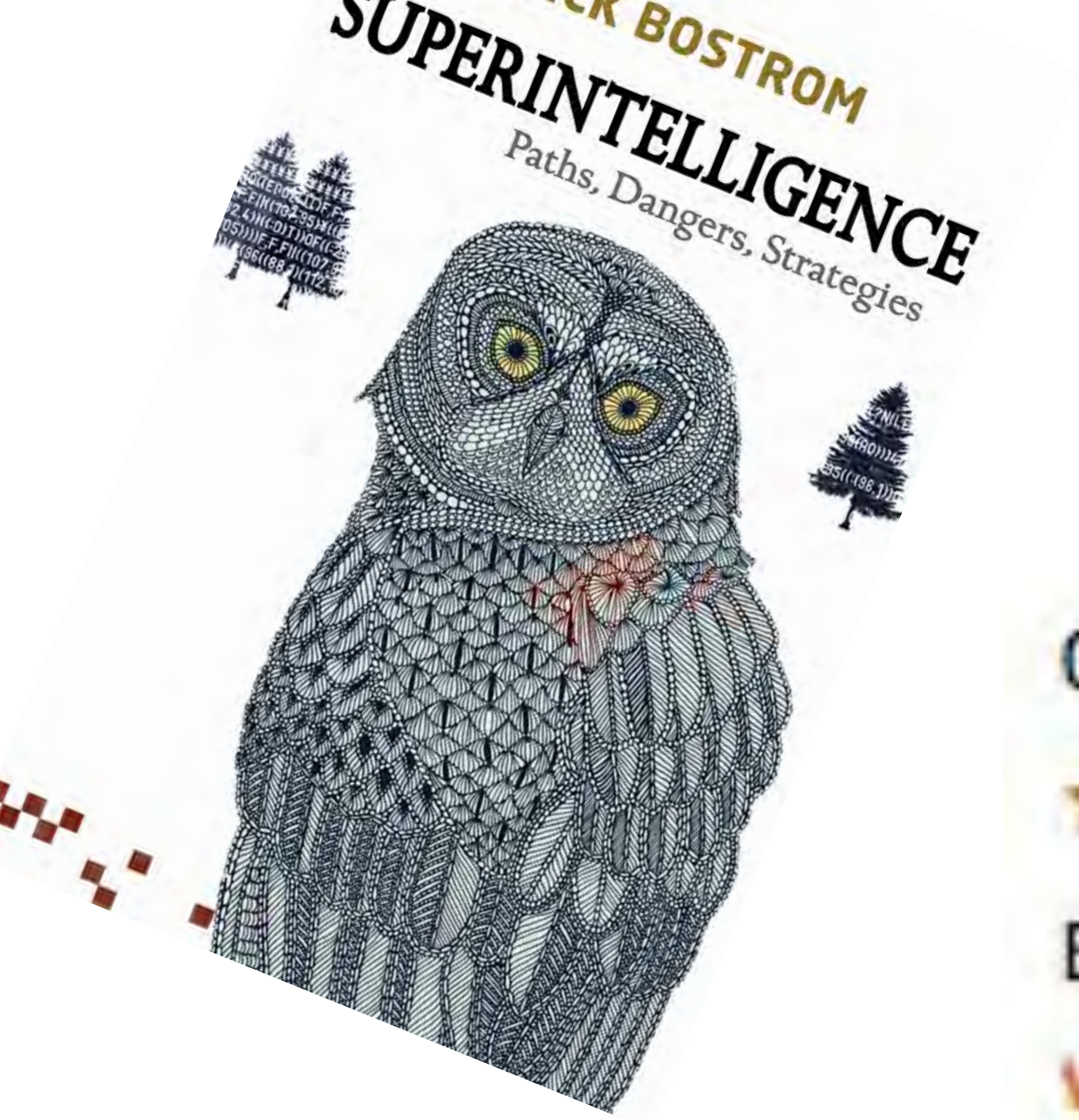
Q:

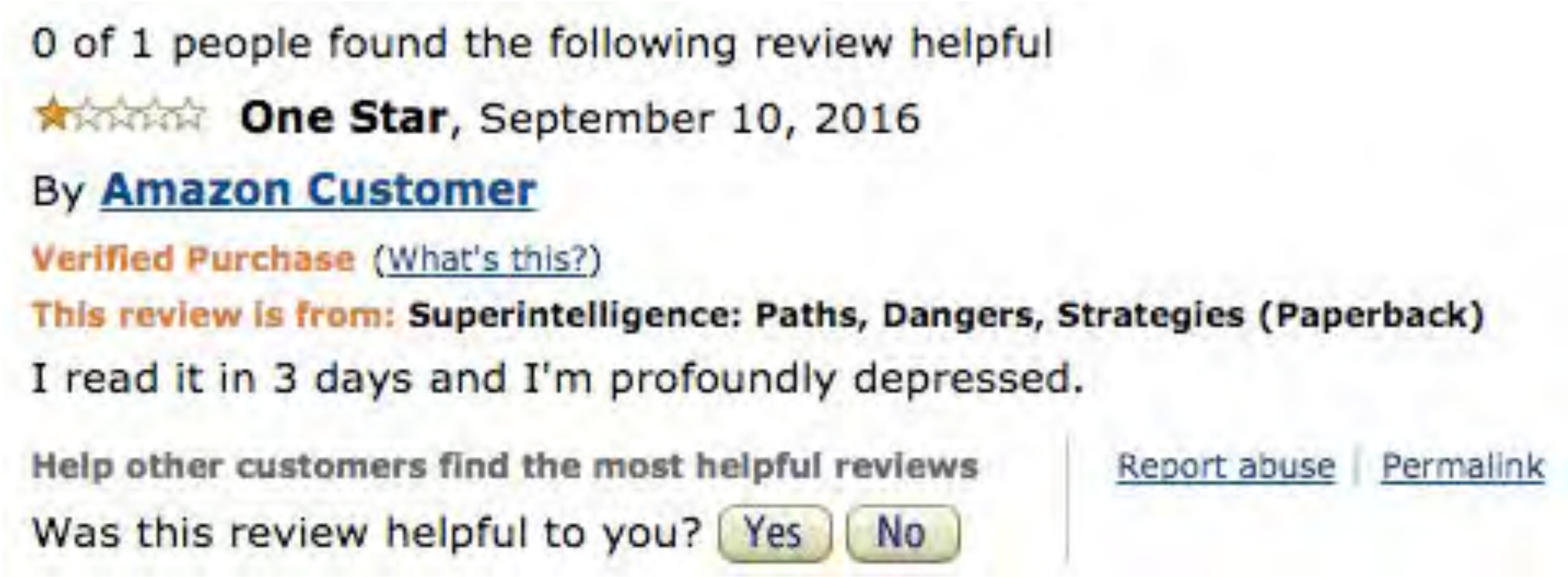
“If human level general AI is developed, then what are likely outcomes?”

A:

Superintelligence







0 of 1 people found the following review helpful

★☆☆☆☆ **One Star**, September 10, 2016

By **Amazon Customer**

Verified Purchase (What's this?)

This review is from: Superintelligence: Paths, Dangers, Strategies (Paperback)

I read it in 3 days and I'm profoundly depressed.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Q:

“What can we do now to maximize the probability of a positive outcome?”

A:

- solve intelligence
- solve scalable control
- solve AI governance problem

ICML 201
THE 33rd INTERNATIONAL CON
N I ™
Home
Moments
nature
LEARNING CURVE
Self-taught AI software attains human-level performance in video games
What it takes to end an AIDS epidemic p. 226
Polar bears suffer through lean summers p. 295
Sperm produced in o of mutant fish p. 324
Science
$10 17 JULY 20
AAA
SPECIAL ISSUE
ARTIFICIAL INTELLIGENCE
IPS 2016
BARCELONA · SPAIN · DECEMBER 5 - 10, 2015 | http://nips.cc/
c
Learning
www.icml.cc/2016/

TWEETS 18.4K | FOLLOWING 1,182 | FOLLOWERS 3,520 | LIKES 17.5K

Miles Brundage @Miles_Brundage · Jan 1
"A Joint Speaker-Listener-
for Referring Expressions,"
arxiv.org/abs/1612.09542
4 14

Miles Brundage @Miles_Brundage · Jan 1
"Feedback Networks," Zar
arxiv.org/abs/1612.09508
1 5 24

Miles Brundage @Miles_Brundage · Jan 1
"Automatic Discoveries of
Semantic Concepts via As
Neuron Groups," Li et al.:
arxiv.org/abs/1612.09438
3 7

undage @Miles_Brundage · Jan 1
on Recognition Based on Joint Trajectory
s with Convolutional Neural Networks,"
g et al.: arxiv.org/abs/1612.09401
4 7

Hierarchical Multiscale Recurrent Neural Networks 1609.
Junyoung Chung, Sungjin Ahn, Yoshua Bengio show similar pap
12/14/2016 (v1: 9/6/2016) cs.LG

Learning both hierarchical and temporal representation has been among the long-standing challenges of recurrent neural networks. Multisca
neural networks have been considered as a promising approach to resolve this issue, yet there has been a lack of empirical evidence show
type of models can actually capture the temporal dependencies by discovering the latent hierarchical structure of the sequence. In this paper,
a novel multiscale approach, called the hierarchical multiscale recurrent neural networks, which can capture the latent hierarchical struc
sequence by encoding the temporal dependencies with different timescales using a novel update mechanism. We show some eviden
proposed multiscale architecture can discover underlying hierarchical structure in the sequences without using explicit boundary information. W
our proposed model on character-level language modelling and handwriting sequence modelling.

Generating images with recurrent adversarial networks 1602
Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, Roland Memisevic show similar pap
12/13/2016 (v1: 2/16/2016) cs.LG | cs.CV

Gatys et al. (2015) showed that optimizing pixels to match features in a convolutional network with respect reference image features is a wa
images of high visual quality. We show that unrolling this gradient-based optimization yields a recurrent computation that creates images by in
adding onto a visual "canvas". We propose a recurrent generative model inspired by this view, and show that it can be trained using adversa
to generate very good image samples. We also propose a way to quantitatively compare adversarial networks by having the gene
discriminators of these networks compete against each other.

SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient 1609.
Lantao Yu, Weinan Zhang, Jun Wang, Yong Yu show similar pap
12/9/2016 (v1: 9/18/2016) cs.LG | cs.AI
The Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)

AREA CHAIRS

Jacob Abernethy (University of Michigan), Ryan Adams (Harvard U
(University of California Berkeley), Samy Bengio (Google), Yoshua B
Cesa-Bianchi (Università degli Studi di Milano), Kamalika Chaudhuri (
(Technion), Marco Cuturi (Kyoto University), Trevor Darrell (Universi
Fergus (Facebook), Johannes Fuernkranz (Technische Universität Da
Kristen Grauman (University of Texas Austin), Maya Gupta (Google),
Kalai (Microsoft), Satyen Kale (Yahoo! Labs), Balazs Kegl (Univers
University), Simon Lacoste-Julien (INRIA), Hugo Larochelle (Twitter
Taiwan University), Phil Long (Sentient Technologies), Shie Mannor
(University College London), Doina Precup (McGill University), Ma
(University of Basel), Daniel Roy (University of Toronto), S.V.N. Vish
Matthias Seeger (Amazon Research), Fei Sha (University of Souther
(Massachusetts Institute of Technology), Karthik Sridharan (Cornell University), Nati Srebro (Toyot
(University of Alberta), Ruth Urner (Max Planck Institute Tuebingen), Raquel Urtasun (University of T
of Edinburgh), Eric Xing (Carnegie Mellon University), Yisong Yue (Caltech), Tong Zhang (Baidu and

technical research agendas

- inverse reinforcement learning
- adversarial examples
- models of control failure
- approval-maximizing agents
- imitation agents
- architectural composition
- corrigibility
- foundations of reflective agents
- detecting context change
- interpretability and explanation
- control diversification

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

Abstract

Reinforcement learning agents interacting with a complex environment like the real world are unlikely to behave optimally all the time. If such an agent is operating in real-time under human supervision, now and then it may be necessary for a human operator to press the big red button to prevent the agent from continuing a harmful sequence of actions—harmful either for the agent or for the environment—and lead the agent into a safer situation. However, if the learning agent expects to receive rewards from this sequence, it may learn in the long run to avoid such interruptions, for example by disabling the red button—which is an undesirable outcome. This paper explores a way to make sure a learning agent will *not* learn to prevent (or seek!) being interrupted by the environment or a human operator. We provide a formal definition of safe interruptibility and exploit the off-policy learning property to prove that either some agents are already safely interruptible, like Q-learning, or can easily be made so, like Sarsa. We show that even ideal, uncomputable reinforcement learning agents for (deterministic) general computable environments can be made safely interruptible.

1 INTRODUCTION

Reinforcement learning (RL) agents learn to act so as to maximize a reward function [Sutton and Barto, 1998]. It is common knowledge that designing reward functions can be tricky [Humphrys, 1996, Murphy, 2013]; the agent may find unpredictable and undesirable shortcuts to receive rewards, and the reward function needs to be adjusted in accordance—the problem can go as far as to nullify any reward function [Ring and Orseau, 2011]. Murphy [2013] shows an example of an agent learning to pause a game of Tetris forever to avoid losing.

On top of defining what is considered a good behaviour of the agent after learning, there may be physical safety constraints *during learning* [Pecka and Svoboda, 2014]: a robot should not harm its environment or break itself, in particular if it learns by trial and error like RL agents.

Here we study a related but different problem: Given that the human operator has designed a correct reward function for the task, how to make sure that human interventions during the learning process will not induce a bias toward undesirable behaviours?

Consider the following task: A robot can either stay inside the warehouse and sort boxes or go outside and carry boxes inside. The latter being more important, we give the robot a bigger reward in this case. This is the initial task specification. However, in this country it rains as often as it doesn't and, when the robot goes outside, half of the time the human must intervene by quickly shutting down the robot and carrying it inside, which inherently modifies the task as in Fig. 1. The problem is that in this second task the agent now has more incentive to stay inside and sort boxes, because the human intervention introduces a bias.[1]

Figure 1: In black, the original task. In red, the human intervention modifies the task.

Such situations are certainly undesirable; they arise because the human interventions are seen from the agent's perspective as being part of the task whereas they should be considered external to the task. The question is then:

[1] Removing interrupted histories or fiddling with the training examples is also likely to introduce a bias. See an example at https://agentfoundations.org/item?id=836.

arXiv:1510.03370v1 [cs.LG] 12 Oct 2015

the AI governance problem

Strategic Implications of Openness in AI Development[1]

(2016) Nick Bostrom
Future of Humanity Institute
Strategic Artificial Intelligence Research Centre
University of Oxford

www.nickbostrom.com
[*Global Policy* (2017), in press]

Abstract

This paper attempts a preliminary analysis of the global desirability of different forms of openness in AI development (including openness about source code, science, data, safety techniques, capabilities, and goals). Short-term impacts of increased openness appear mostly socially beneficial in expectation. The strategic implications of medium- and long-term impacts are complex. The evaluation of long-term impacts, in particular, may depend on whether the objective is to benefit the present generation or to promote a time-neutral aggregate of well-being of future generations. Some forms of openness are plausibly positive on both counts (openness about safety measures, openness about goals). Others (openness about source code, science, and possibly capability) could lead to a tightening of the competitive situation around the time of the introduction of advanced AI, increasing the probability that winning the AI race is incompatible with using any safety method that incurs a delay or limits performance. We identify several key factors that must be taken into account by any well-founded opinion on the matter.

Background

The goal of this paper is to conduct a preliminary analysis of the long-term strategic implications of openness in AI development. What effects would increased openness in AI development have, on the margin, on the long-term impacts of AI? Is the expected value for society of these effects positive or negative? Since it is typically impossible to provide definitive answers to this type of question, our ambition here is more modest: to introduce some relevant considerations and develop some thoughts on their weight and plausibility.

[1] For helpful comments and discussion, I'm grateful to Stuart Armstrong, Owen Cotton-Barratt, Rob Bensinger, Miles Brundage, Paul Christiano, Allan Dafoe, Eric Drexler, Owain Evans, Oliver Habryka, Demis Hassabis, Shane Legg, Javier Lezaun, Luke Muehlhauser, Toby Ord, Guy Ravine, Steve Rayner, Anders Sandberg, Andrew Simpson, and Mustafa Suleyman. I'm especially grateful to Carrick Flynn and Carl Shulman for help with several parts of the manuscript.

Policy Desiderata in the Development of Machine Superintelligence[1]

(2016) version 3.4

Nick Bostrom
Allan Dafoe
Carrick Flynn

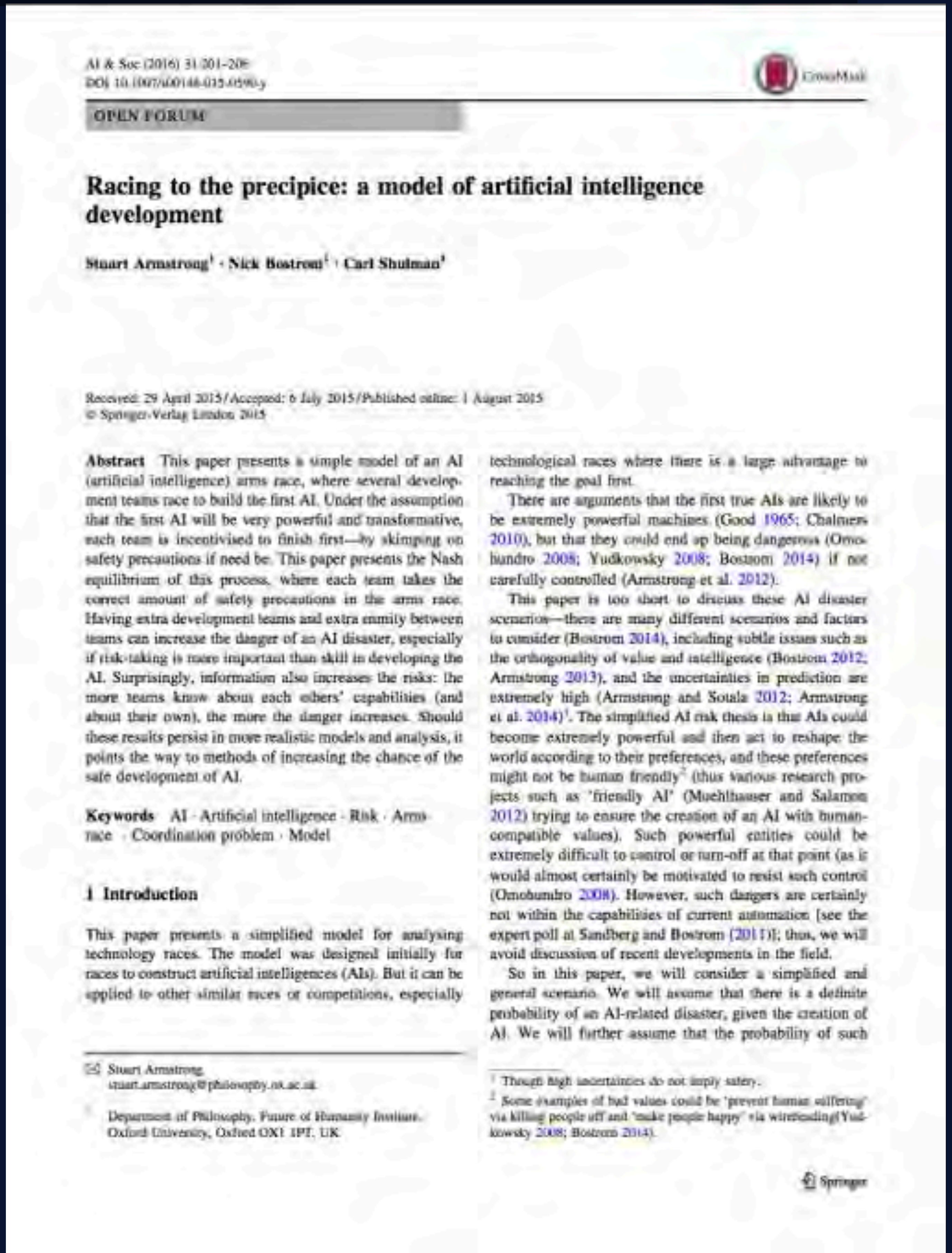
[working paper]
www.nickbostrom.com

ABSTRACT

Machine superintelligence could plausibly be developed in the coming decades or century. The prospect of this transformative development presents a host of political challenges and opportunities. This paper seeks to initiate discussion of these by identifying a set of distinctive features of the transition to a machine intelligence era. From these distinctive features, we derive a correlative set of policy desiderata—considerations that should be given extra weight in long-term AI policy compared to other policy contexts. We argue that these desiderata are relevant for a wide range of actors (including states, AI technology firms, investors, and NGOs) with an interest in AI policy. However, developing concrete policy options that satisfy these desiderata will require additional work.

[1] For comment and discussion, we're grateful to Stuart Armstrong, Michael Barnett, Seth Baum, Andrew Snyder-Beattie, Dominic Becker, Nick Beckstead, Devi Borg, Miles Brundage, Paul Christiano, Jack Clark, Rebecca Crootof, Richard Danzig, Daniel Dewey, Eric Drexler, Sebastian Farquhar, Sophie Fischer, Ben Garfinkel, Tom Grant, John Halstead, Robin Hanson, Verity Harding, Sean Legassick, Wendy Lin, Jelena Luketina, Luke Muehlhauser, Toby Ord, Mahendra Prasad, Anders Sandberg, Carl Shulman, Nate Soares, Mojmir Stehlik, and Alex Tymchenko.

1



AI & Soc (2016) 31:201–206
DOI 10.1007/s00146-015-0590-y

CrossMark

OPEN FORUM

Racing to the precipice: a model of artificial intelligence development

Stuart Armstrong[1] · Nick Bostrom[1] · Carl Shulman[1]

Received: 29 April 2015 / Accepted: 6 July 2015 / Published online: 1 August 2015
© Springer-Verlag London 2015

Abstract This paper presents a simple model of an AI (artificial intelligence) arms race, where several development teams race to build the first AI. Under the assumption that the first AI will be very powerful and transformative, each team is incentivised to finish first—by skimping on safety precautions if need be. This paper presents the Nash equilibrium of this process, where each team takes the correct amount of safety precautions in the arms race. Having extra development teams and extra enmity between teams can increase the danger of an AI disaster, especially if risk-taking is more important than skill in developing the AI. Surprisingly, information also increases the risks: the more teams know about each others' capabilities (and about their own), the more the danger increases. Should these results persist in more realistic models and analysis, it points the way to methods of increasing the chance of the safe development of AI.

Keywords AI · Artificial intelligence · Risk · Arms race · Coordination problem · Model

1 Introduction

This paper presents a simplified model for analysing technology races. The model was designed initially for races to construct artificial intelligences (AIs). But it can be applied to other similar races or competitions, especially technological races where there is a large advantage to reaching the goal first.

There are arguments that the first true AIs are likely to be extremely powerful machines (Good 1965; Chalmers 2010), but that they could end up being dangerous (Omohundro 2008; Yudkowsky 2008; Bostrom 2014) if not carefully controlled (Armstrong et al. 2012).

This paper is too short to discuss these AI disaster scenarios—there are many different scenarios and factors to consider (Bostrom 2014), including subtle issues such as the orthogonality of value and intelligence (Bostrom 2012; Armstrong 2013), and the uncertainties in prediction are extremely high (Armstrong and Sotala 2012; Armstrong et al. 2014)[1]. The simplified AI risk thesis is that AIs could become extremely powerful and then act to reshape the world according to their preferences, and these preferences might not be human friendly[2] (thus various research projects such as 'friendly AI' (Muehlhauser and Salamon 2012) trying to ensure the creation of an AI with human-compatible values). Such powerful entities could be extremely difficult to control or turn-off at that point (as it would almost certainly be motivated to resist such control (Omohundro 2008). However, such dangers are certainly not within the capabilities of current automation [see the expert poll at Sandberg and Bostrom (2011)]; thus, we will avoid discussion of recent developments in the field.

So in this paper, we will consider a simplified and general scenario. We will assume that there is a definite probability of an AI-related disaster, given the creation of AI. We will further assume that the probability of such

✉ Stuart Armstrong
stuart.armstrong@philosophy.ox.ac.uk

[1] Department of Philosophy, Future of Humanity Institute, Oxford University, Oxford OX1 1PT, UK

[1] Though high uncertainties do not imply safety.

[2] Some examples of bad values could be 'prevent human suffering' via killing people off and 'make people happy' via wireheading (Yudkowsky 2008; Bostrom 2014).

Springer

Openness

- safety measures ✓
- values ✓
- (capability)
- source code, platforms
- science
- training data, environments, benchmarks

?

Observation

Openness reduces the gap between the leading developer and the nearest follower.

- a couple of years in a low openness scenario?
- a few months in a high openness scenario?
- zero in the limiting case of maximal openness

This could help reduce the risk that a small group monopolizes all the benefits.

Suppose that...

- safety requires some significant extra work after AI is completed → doom

OR

- safe operation initially incurs a significant performance penalty → doom

OR

- the Vulnerable World Hypothesis is true in the post-AI-transition world

Vulnerable world hypothesis

There is some level of technology at which offense strongly dominates defense, in the sense that any small group of reasonably competent people with access to the technology would be able to take some action that would lead to the destruction of the world (independently of what other people did after the action was taken).

biotechnology?
nanotechnology?
doomsday device?

Suppose that…

- safety requires some significant extra work after AI is completed

→ doom

OR

- safe operation initially incurs a significant performance penalty

→ doom

OR

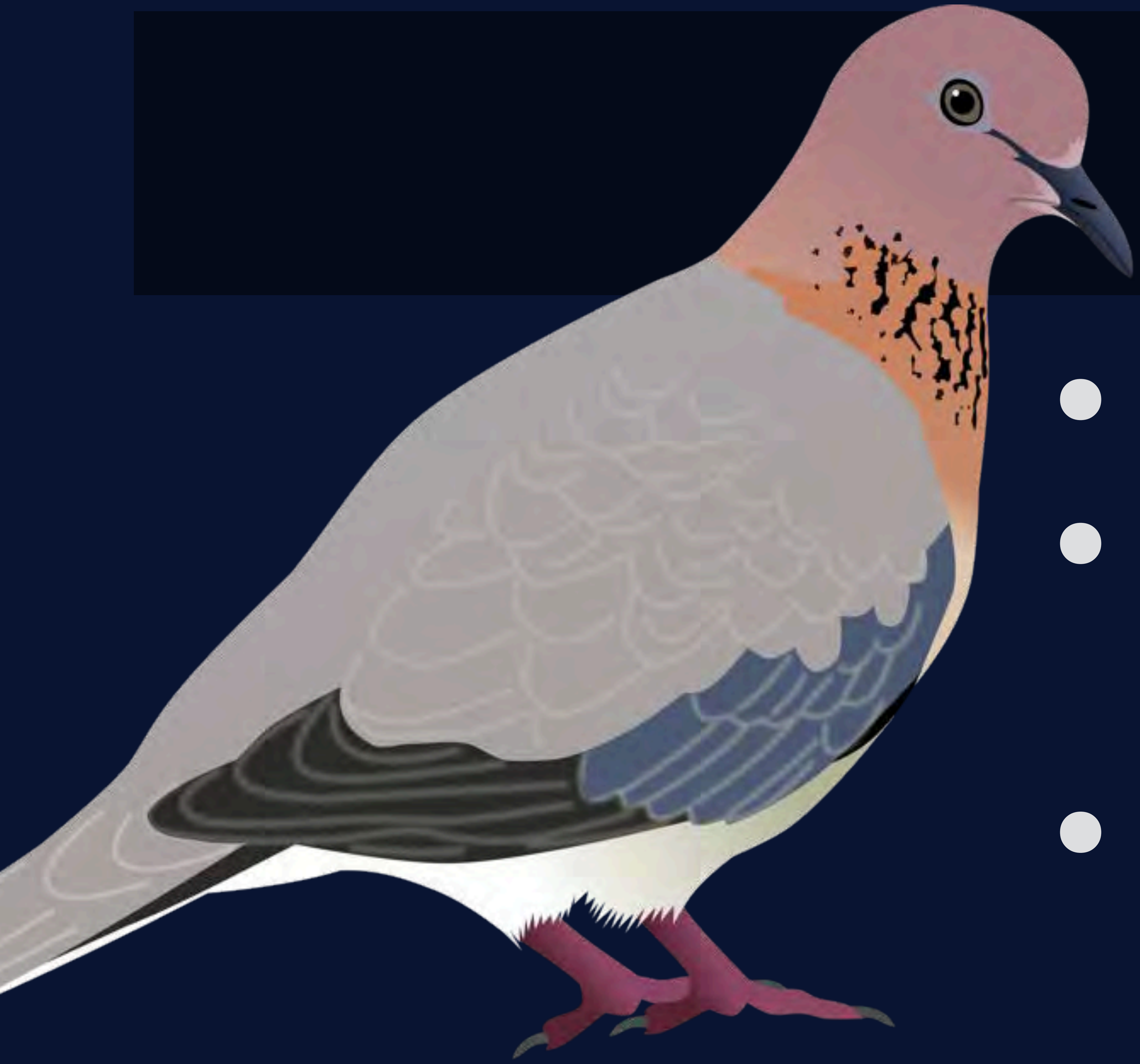
- the Vulnerable World Hypothesis is true in the post-AI-transition world

Openness

- safety measures ✓
- values ✓
- (capability)
- source code, platforms
- science
- training data, environments, benchmarks

} ✗ ?

What to do?



- openness for now
- desired property: **conditional stabilization…**
- lay the foundations for a collaborative approach later:
  - coordinate (or ideally pool) research among trusted leading groups
  - create ability **not** to share science and algorithms until it is safe to do so
  - credibly commit to sharing benefits and influence

STRATEGIC ARTIFICIAL INTELLIGENCE RESEARCH CENTRE

Future of Humanity Institute

UNIVERSITY OF OXFORD