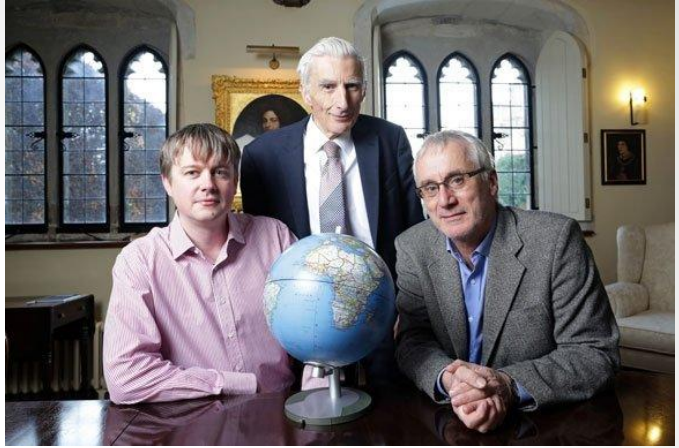
Organizations focusing on existential risks

Victoria Krakovna, Harvard

Existing organizations

- Centre for the Study of Existential Risk (CSER)
- Future of Humanity Institute (FHI)
- Machine Intelligence Research Institute (MIRI)
- Global Catastrophic Risk Institute (GCRI)
- **Less focused:** Union of Concerned Scientists, Federation of American Scientists, Foresight Institute, JASON defense advisory group
- **Less active:** Center for Responsible Nanotechnology, Lifeboat Foundation

CSER



- Cambridge, UK
- Founded in 2012
- Founders: Jaan Tallinn, Martin Rees, Huw Price
- First public lecture on Feb 26, 2014
- Research areas:

synthetic biology	AI
nanotechnology	global warming

systemic risks and fragile networks

CSER's plan

- Engage with top researchers in the relevant fields
- Act as a bridge between experts from different fields, from academia and industry
- Organize research workshops and conferences

FHI

Future of Humanity Institute
UNIVERSITY OF OXFORD

- Oxford, UK
- Founded in 2005 by Nick Bostrom
- Major researchers: Nick Bostrom, Stuart Armstrong, Anders Sandberg, Seán Ó hÉigeartaigh, etc
- Offshoot: Oxford Martin Programme on the Impacts of Future Technology (launched in 2011)

Future of Humanity Institute
UNIVERSITY OF OXFORD

FHI's research areas

- Existential risks
  - Papers: "Existential Risk Prevention as Global Priority", "Methodological Challenges for Risks with Low Probability and High Stakes"
- Future technology forecasting
  - Paper: "Whole Brain Emulation: A Roadmap"
- Applied epistemology
  - Paper: "Anthropic Bias: Observation Selection Effects in Science and Philosophy"

MIRI

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

- Berkeley, CA
- Formerly SIAI, founded by Eliezer Yudkowsky in 2000, currently led by Luke Muehlhauser
- Focused on building a provably safe Strong AGI (“Friendly” AI)
- Major researchers: Eliezer Yudkowsky, Benja Fallenstein, Marcello Herreshoff, Paul Christiano, etc

MIRI's research directions

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

- Self-reflective agents
  - Paper: "Tiling Agents for Self-Modifying AI"
- Decision theory
  - Paper: "Robust Cooperation on the Prisoner's Dilemma"
- Value functions
  - Paper: "Avoiding Unintended AI Behaviors"
- Forecasting
  - "How We're Predicting AI—or Failing To"
  - "Intelligence Explosion: Evidence and Import"

GCRI

Global Catastrophic Risk INSTITUTE

- founded in 2011 by Seth Baum and Tony Barrett
- Geographically decentralized, mostly based in NYC, SF, Washington
- Major researchers: Grant Wilson, Timothy Maher, Jacob Haqq-Misra
- connections to Society for Risk Analysis (SRA)

GCRI research areas

- Specific GCR's
  - Paper: "Minimizing global catastrophic and existential risks from **emerging technologies** through international law."
  - Paper: "Analyzing and reducing the risks of inadvertent **nuclear war** between the United States and Russia."
- Cross-cutting topics
  - Paper: "The **ethics** of global catastrophic risk from dual-use bioengineering."
- Synthesis (risk interaction)
  - Papers: "When global catastrophes collide: The climate engineering double catastrophe", "**Double catastrophe**: Intermittent stratospheric geoengineering induced by societal collapse."

Other organizations

- Union of Concerned Scientists,

  Federation of American Scientists
  - Focused on global warming and nuclear risks
- JASON Defense Advisory Group
  - Papers: “Human Performance Modification”, “Opportunities at the Intersection of Nanoscience, Biology and Computation”, “Rare Events”
- Foresight Institute
  - Annual conferences, focused on advancing nanotechnology
- Lifeboat Foundation
  - Recent papers are few and irrelevant

Where do we fit in?

- Create an East Coast hub focusing on x-risk
- Bring together local intellectuals interested in these essential questions (Harvard, MIT, etc)
- Cooperate with sister organizations
- Broader focus than just research (advocacy, conferences, etc.)