About the Future of Humanity Institute

FHI has originated or played a pioneering role in developing many of the key concepts that shape current thinking about humanity’s future. These include: simulation argument, existential risk, nanotechnology, information hazards, strategy and analysis related to machine superintelligence, astronomical waste, the ethics of digital minds, crucial considerations, observation selection effects in cosmology and other contexts of self-locating belief, prediction markets, infinitarian paralysis, brain emulation scenarios, human enhancement, the unilateralist’s curse, the parliamentary model of decision making under normative uncertainty, the vulnerable world hypothesis, and many others.

Contents

Future of Humanity Institute (FHI)

Description:

The Future of Humanity Institute is a unique world-leading research centre that works on big picture questions for human civilisation and explores what can be done now to ensure a flourishing long-term future. Its multidisciplinary research team includes several of the world's most brilliant and famous minds working in this area. Its work spans the disciplines of mathematics, philosophy, computer science, engineering, ethics, economics, and political science.

Stakeholder(s):

University of Oxford

Nick Bostrom :
Director ~ Nick Bostrom is a Swedish-born philosopher with a background in theoretical physics, computational neuroscience, logic, and artificial intelligence, as well as philosophy. He is a Professor at Oxford University, where he heads the Future of Humanity Institute as its founding director. Bostrom is the most-cited professional philosopher under the age of 50. He is the author of some 200 publications, including Anthropic Bias (2002), Global Catastrophic Risks (2008), Human Enhancement (2009), and Superintelligence: Paths, Dangers, Strategies (2014), a New York Times bestseller which helped spark a global conversation about the future of AI. He has also published a series of influential papers, including ones that introduced the simulation argument (2003) and the concept of existential risk (2002). His academic work has been translated into more than 30 languages. He is a repeat main TED speaker and has been interviewed more than 1,000 times by various media. He has been on Foreign Policy's Top 100 Global Thinkers list twice and was included in Prospect's World Thinkers list, the youngest person in the top 15. As a graduate student he dabbled in stand-up comedy on the London circuit, but he has since reconnected with the heavy gloom of his Swedish roots.

FHI Research Scholars Programme Participants

Lukas Finnveden :
Research Scholar

Spencer Becker-Kahn :
Senior Research Scholar

Angela Aristizabal :
Research Scholar

Luca Righetti :
Research Scholar

Fin Moorhouse :
Research Scholar

Damon Binder :
Senior Research Scholar

DPhil Scholars & Affiliates

Jan Brauner :
DPhil Scholar

Isaac Friend :
DPhil Scholar

Hannah Klim :
DPhil Scholar

Carla Zoe Cremer :
DPhil Scholar

FHI Research Associates

Nick Beckstead :
Open Philanthropy Project

Hilary Greaves :
Global Priorities Institute

Paul Christiano :
Alignment Research Center

William MacAskill :
Global Priorities Institute

Robin Hanson :
George Mason University

Helen Toner :
Center for Security and Emerging Technology

Jan Leike :
OpenAI

Roger Grosse :
Vector Institute, University of Toronto

External DPhil Supervisors

Michael Bonsall :
Professor of Mathematical Biology, University of Oxford

Michael Osborne :
Professor of Machine Learning, University of Oxford

Duncan Snidal :
Professor of International Relations, University of Oxford

Karolina Milewicz :
Associate Professor of International Relations, University of Oxford

Vision

A flourishing long-term future

Mission

To bring the tools of mathematics, science, and philosophy to bear on big-picture questions about humanity and its prospects.

Values

Mathematics

Science

Philosophy

Humanity

1. Humanity

Research topics related to humanity’s future

FHI has individual researchers working across many topics related to humanity’s future.

2. Macrostrategy

Analyze the connections between long-term outcomes and present actions.

Stakeholder(s)

Macrostrategy Research Group

Anders Sandberg :
Senior Research Fellow

Matthew van der Merwe :
Research Assistant to the Director

Toby Ord :
Senior Research Fellow

David 'davidad' Dalrymple :
Research Fellow

Eric Drexler :
Senior Research Fellow

Jan Kulveit :
Research Fellow

Ben Garfinkel :
Research Fellow

Goodwin Gibbins :
Research Fellow

Macrostrategy: How long-term outcomes for humanity are connected to present-day actions; global priorities; crucial considerations that may reorient our civilizational scheme of values or objectives.

2.1. Priorities

Consider global priorities

2.2. Values & Objectives

Address considerations that may reorient our civilizational scheme of values or objectives

3. Artificial Intelligence

3.1. Governance

Consider how humanity can best navigate the transition to advanced AI systems

Governance of Artificial Intelligence: The governance concerns of how humanity can best navigate the transition to advanced AI systems; how geopolitics, governance structures, and strategic trends shape the development or deployment of machine intelligence.

Stakeholder(s):

AI Governance Research Group

Ben Garfinkel :
Research Fellow

3.2. Geopolitics, Structures & Trends

Consider how geopolitics, governance structures, and strategic trends shape the development or deployment of machine intelligence.

3.3. Safety & Values

Identify techniques for building artificially intelligent systems that are scalably safe and aligned with human values

AI Safety: Techniques for building artificially intelligent systems that are scalably safe or aligned with human values (in close collaboration with labs such as DeepMind, OpenAI, and CHAI).

Stakeholder(s):

AI Safety Research Group

Ryan Carey :
DPhil Scholar

Jennifer Lin :
Senior Research Fellow

Chris van Merwijk :
Researcher

4. Biosecurity

Consider how to make the world more secure against biological risks

Stakeholder(s)

Biosecurity Research Group

Cassidy Nelson :
Acting Co-Lead

Jonas Sandbrink :
Researcher

Gregory Lewis :
Acting Co-Lead

Joshua Monrad :
Researcher

Piers Millett :
Senior Research Fellow

Michael Montague :
Research Associate

James Wagstaff :
Research Fellow

Biosecurity: How to make the world more secure against (both natural and human-made) catastrophic biological risks; how to ensure that capabilities created by advances in synthetic biology are handled well.

5. Digital Minds

Consider philosophy of mind and AI ethics

Stakeholder(s)

Digital Minds Research Group

Robert Long :
Research Fellow

Patrick Butlin :
Research Fellow

Carl Shulman :
Research Associate

Digital Minds: Philosophy of mind and AI ethics, focusing on questions concerning which computations are conscious and which digital minds have what kinds of moral status, and what political systems would enable a harmonious coexistence of biological and nonbiological minds.

5.1. Consciousness

Consider which computations are conscious

5.2. Moral Status

Consider which digital minds have what kinds of moral status

5.3. Political Systems

Consider what political systems would enable a harmonious coexistence of biological and nonbiological minds

6. Other Issues

Other areas in which we are active and are interested in expanding include (but are not limited to) the following:

6.1. Uncertainties

Consider how deep uncertainties affect decision making

Philosophical Foundations: When and how might deep uncertainties related to anthropics, infinite ethics, decision theory, computationalism, cluelessness, and value theory affect decisions we might make today? Can we resolve any of these uncertainties?

6.2. Risk

Identify and characterise risks to humanity

Existential Risk: Identification and characterisation of risks to humanity; improving conceptual tools for understanding and analysing these risks.

6.3. Technological Maturity

Consider opportunities available to technologically mature civilizations

Grand Futures: Questions related to the Fermi paradox, cosmological modelling of the opportunities available to technologically mature civilizations, implications of multiverse theories, the ultimate limits to technological advancement, counterfactual histories or evolutionary trajectories, new physics.

6.4. Cooperation

Investigate structures that facilitate future cooperation at different scales

Cooperative Principles and Institutions: Theoretical investigations into structures that facilitate future cooperation at different scales and search for levers to increase the chances of cooperative equilibria, e.g. with respect to rival AI developers, humans and digital minds, or among technologically mature civilizations.

6.5. Technology & Wisdom

Consider how to enable society to act with greater wisdom both in developing and deploying new capabilities

Technology and Wisdom: What constitutes wisdom in choosing which new technological paths to pursue? Are there structures which enable society to act with greater wisdom both in making choices about what to develop and when or how to deploy new capabilities?

6.6. Information Systems

Consider how to design global information systems to mitigate epistemic dysfunctions

Sociotechnical Information Systems: Questions concerning the role of surveillance in preventing existential risks and how to design global information systems (e.g. recommender systems, social networks, peer review, discussion norms, prediction markets, futarchy) to mitigate epistemic dysfunctions.

6.7. Malevolence

Consider how to reduce risk from malevolent humans

Reducing Risk from Malevolent Humans: Defining and operationalizing personality traits of potential concern (e.g. sadism, psychopathy, etc.) or promise (e.g. compassion, wisdom, integrity), especially ones relevant to existential risk; evaluating possible intervention strategies (e.g. cultural and biological mechanisms for minimising malevolence; personnel screening tools; shaping incentives in key situations).

6.8. AI Systems

Understand the scaling properties and limitations of current AI systems

Concepts, Capabilities, and Trends in AI: Understanding the scaling properties and limitations of current AI systems; clarifying concepts used to analyze machine learning models and RL agents; assessing the latest breakthroughs and their potential contribution towards AGI; projecting trends in hardware cost and performance.

6.9. Space

Consider legal system for long-term space development

Space Law: What would be an ideal legal system for long-term space development, and what opportunities exist for adjusting existing treaties and norms?

6.10. Nanotechnology

Analyze roadmaps to atomically precise manufacturing and related technologies

Nanotechnology: Analyzing roadmaps to atomically precise manufacturing and related technologies, and potential impacts and strategic implications of progress in these areas.

Administrative Information

Start Date:
End Date:

Publication Date: **2022-08-24**
Source: https://www.fhi.ox.ac.uk/about-fhi/

Submitter:

Given Name: Owen
Surname: Ambur
Email: Owen.Ambur@verizon.net
Phone:

PDF formatted using TopLeaf XML publisher
www.turnkey.com.au