Highlights from Asilomar workshop on Beneficial AI

Viktoriya Krakovna, FLI / DeepMind

AI safety research areas

Many challenges associated with increasing AI capabilities:

- Value learning
- Robust self-modification
- Anomaly detection
- Governance and policy
- ... and more

Good news: Awesome people working on these problems **right now!**

Value learning

Question: How to specify complicated human values and ethics to AI systems?

Value learning by human feedback

Stuart Russell: Teach the agent by demonstrating human actions (cooperative inverse reinforcement learning).

Owain Evans: Human actions are often inconsistent and suboptimal. Modify inverse reinforcement learning to account for human biases.

Paul Christiano: Use semi-supervised learning to decrease reliance on human feedback (scalable AI control).

Value learning by building in morality

Francesca Rossi: Specify ethical laws through constraints.

Vincent Conitzer: Find patterns in human ethical decisions, and build those features into AI systems.

Adrian Weller: Can we make human moral concepts more precise and consistent?

Robust self-modification

Question: How can AI systems modify themselves while retaining their safety properties?

Robust self-modification

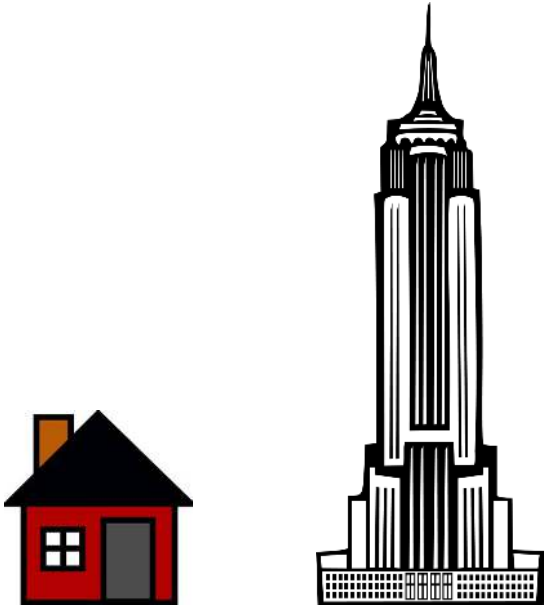
Ramana Kumar: Implement a formal verification model to study the challenges of self-referential reasoning.

Bas Steunebrink: Bounded recursive self-modification: make small modifications and test them empirically.

Anomaly detection

Question: How can AI systems recognize when they are in an unfamiliar setting and generalize from their past experiences?

"There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know."



Anomaly detection

Percy Liang: You can make good predictions even without assuming where your test data comes from.

Tom Dietterich: Use a monitoring algorithm to detect when the original algorithm is extrapolating.

Fuxin Li: Never do extrapolation! Test whether a data point is normal or adversarial.

Brian Ziebart: Be pessimistic! What is the worst case for predictive data that still matches the previous observations?

Governance and policy

Question: How can we help policymakers manage the societal impacts of AI?

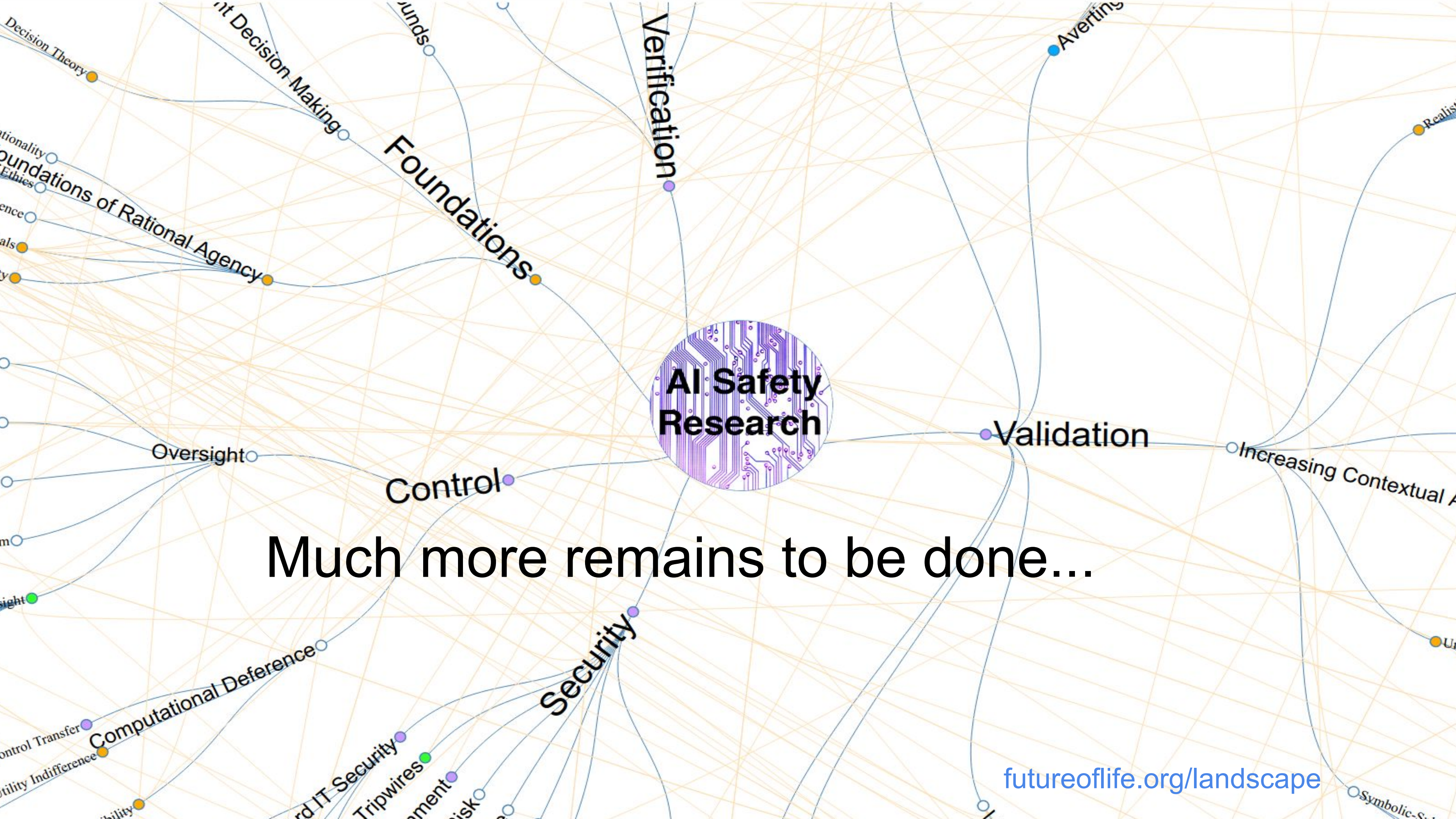
Governance and policy

Heather Roff: Define the concept of meaningful human control of autonomous weapons at tactical, operational, and strategic levels.

Peter Asaro: Who is responsible for the actions of autonomous weapons? Define what we mean by autonomy, agency, and liability.

Moshe Vardi: Organize a multidisciplinary summit on job automation.

Nick Bostrom: Derive policy desiderata for transition to machine intelligence era: efficiency, coordination, common good



Much more remains to be done...

futureoflife.org/landscape

Current AI safety research teams

Academia:

CENTRE FOR THE STUDY OF EXISTENTIAL RISK
UNIVERSITY OF CAMBRIDGE

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Future of Humanity Institute
UNIVERSITY OF OXFORD

UC Berkeley Center for Human-Compatible AI

FLI grantees

Independent:

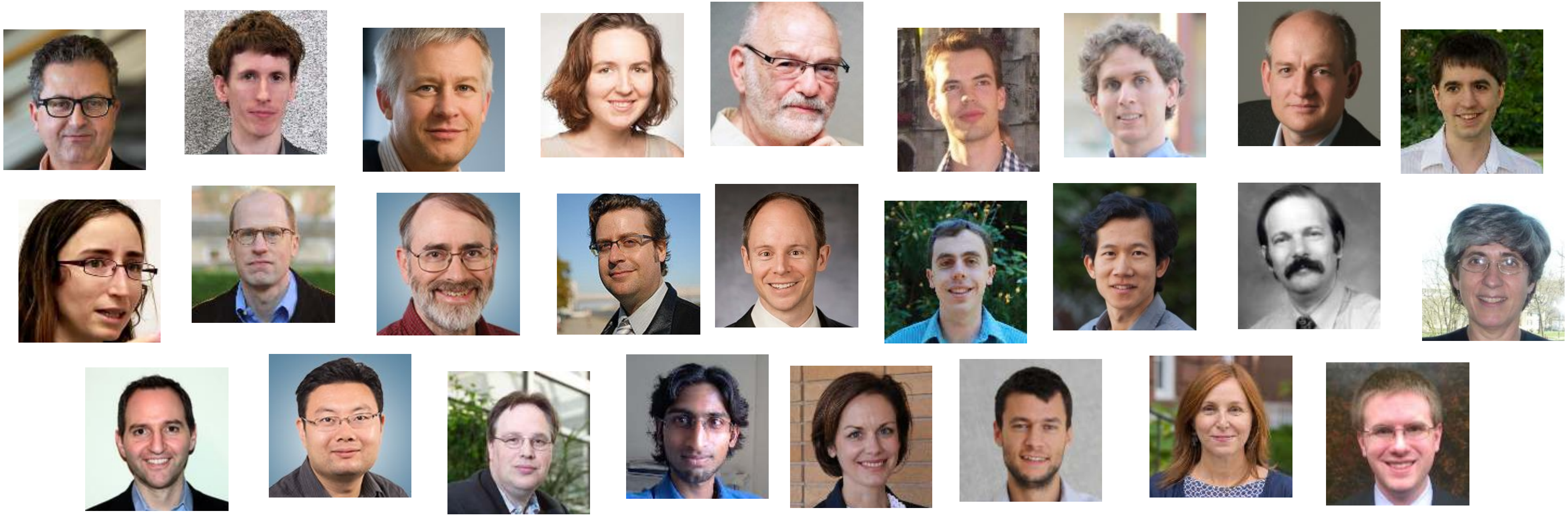
MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

Industry:

DeepMind

OpenAI

Have a chat with the FLI researchers!



futureoflife.org/ai-safety-research

KEEP
CALM

AND

WORK ON
AI SAFETY