Technology is giving life the potential to flourish like never before...

...or to self-destruct. Let's make a difference!

future of life INSTITUTE

Annual Report

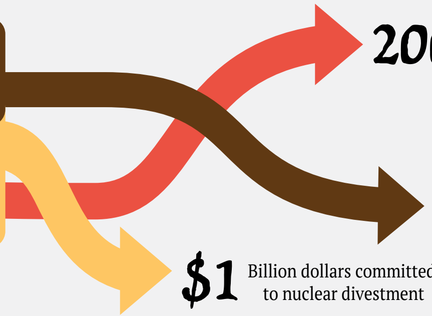
Making a Difference

2016

2016 By the Numbers

01 Nuclear Conference

We launched our nuclear divestment campaign at MIT'S Nuclear Conference.



200 Participants

18 Speakers and panelists

$1 Billion dollars committed to nuclear divestment

43 AI Safety Events

This is an increase of 36 AI safety events from the previous year.

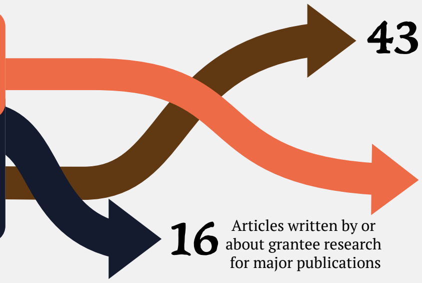
17 Conferences attended by FLI members

15 Invitations to speak at events

4 Events co-organized to highlight AI safety

37 Grant Recipients

Our grant recipients are already hard at work to keep AI beneficial.

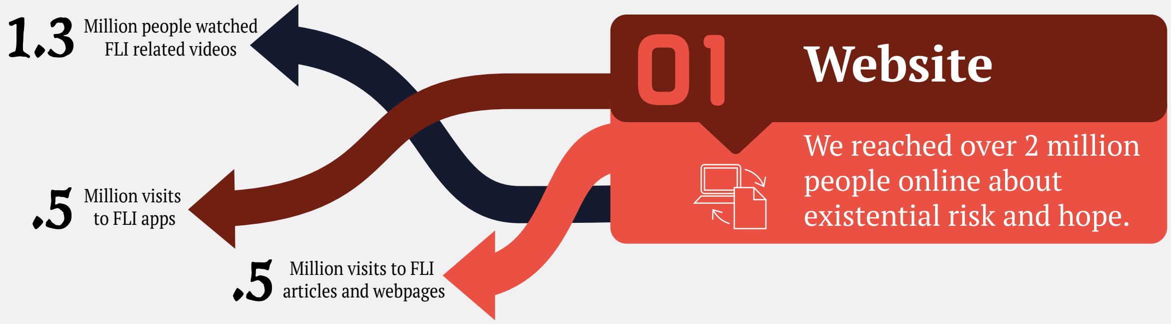


43 Published, peer-reviewed papers

87 Workshops grantees organized and/or participated in

16 Articles written by or about grantee research for major publications

2016 By the Numbers, Cont.



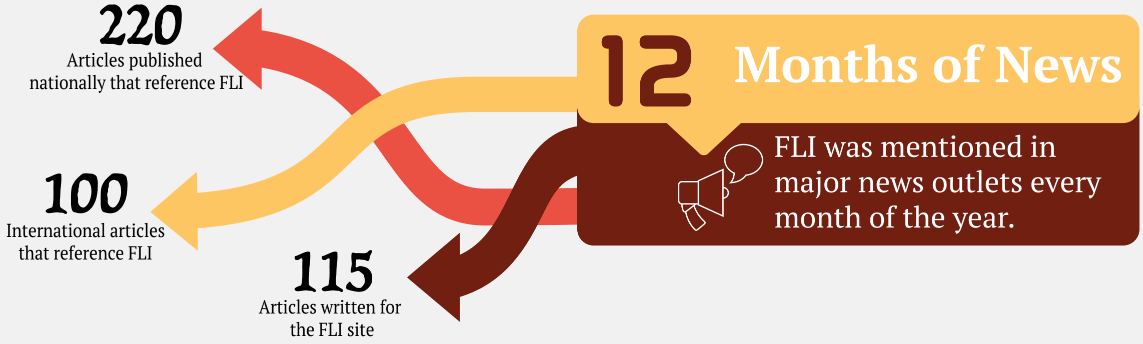
01 Website

We reached over 2 million people online about existential risk and hope.

1.3 Million people watched FLI related videos

.5 Million visits to FLI apps

.5 Million visits to FLI articles and webpages



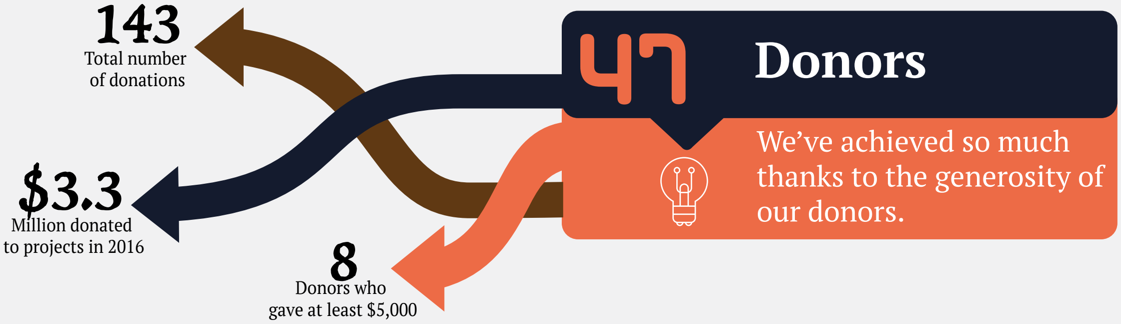
12 Months of News

FLI was mentioned in major news outlets every month of the year.

220 Articles published nationally that reference FLI

100 International articles that reference FLI

115 Articles written for the FLI site



47 Donors

We've achieved so much thanks to the generosity of our donors.

143 Total number of donations

$3.3 Million donated to projects in 2016

8 Donors who gave at least $5,000

Message from the President

It's been a great honor for me to get to work with such a talented and idealistic team at our institute to ensure that tomorrow's most powerful technologies have a positive impact on humanity. With less powerful technologies such as fire, we learned from mistakes, but with more powerful technologies such as nuclear weapons and future strong artificial intelligence, planning ahead is a better strategy. We worked hard in 2016, supporting research and other efforts to avoid problems in the first place.

We are delighted to have helped the beneficial-AI movement go mainstream and to have launched the first-ever global research program aimed at keeping AI safe, awarding $7.1M to 37 talented teams that have already published scores of scientific papers with interesting results. Extensive groundwork in 2016 helped make our Asilomar AI Conference a great success, articulating key steps toward beneficial AI and building community consensus around them. We also made major efforts in 2016 toward reducing the risk of nuclear war: we developed popular online educational tools, helped run a large and successful conference at MIT, and helped stigmatize the production of destabilizing new nuclear weapons by launching a divestment campaign that accomplished a billion dollar commitment. Our site futureoflife.org brought educational materials and news to ever more people, doubling its viewership in 2016.

As we look ahead to 2017, we see great opportunities to expand these efforts. To ensure that AI isn't merely robust but also improves global well-being, the burgeoning beneficial-AI community needs to extend into the social sciences. In 2017, the United Nations will explore curtailing arms races in both nuclear weapons and lethal autonomous weapons, and we will support these efforts by engaging scientists and experts and by making the key facts widely accessible. We aim to scale up our fledgling outreach efforts to give the quest for the long-term success of life the prominence it deserves. Technology is giving life the potential to flourish like never before – let's seize this opportunity together!

-Max Tegmark

Some of the FLI grant winners at the AAAI conference.

Major Accomplishments of 2016

Former Defense Secretary William Perry and Mayor Denise Simmons were among the speakers at the nuclear conference.

Nuclear Conference & Divestment

In 2016, FLI launched various efforts to reduce the threat of nuclear weapons. Among them was our divestment campaign, which we officially launched at the April 2 Nuclear Conference at MIT. It was there that Mayor Denise Simmons of Cambridge, Massachusetts announced that, with FLI's help, the city council had unanimously voted to divest their $1 billion pension fund from nuclear-weapons producing companies.

The meeting attracted over 200 attendees, and included such speakers and panelists as Former Defense Secretary William Perry, Nobel laureate Frank Wilczek, Ploughshares Fund President Joe Cirincione, and many others.

At that meeting, we revealed our MinutePhysics video, *Why You Should Care About Nukes*, which has been viewed by over 1.3 million people to date. We also rolled out a nuclear divestment app, which is a project we look forward to doing more with in 2017.

Throughout 2016, we continued to build collaborations with various groups aligned with our goal to reduce the risk of nuclear weapons, and we look forward to working with them further in 2017, as we expand our nuclear divestment campaign and our outreach efforts.

Continued Impact of the Open Letters

The 2015 open letters on Artificial Intelligence and Autonomous Weapons continued to have an impact in 2016. The media referenced them often, and they were cited in the White House AI reports, helping to cement AI safety as an important and much more mainstream research topic. The autonomous weapons letter also contributed to the recent United Nations decision to establish a Group of Governmental Experts, which will meet in 2017 and report back to the Convention on Certain Conventional Weapon's annual meeting in November.

AI Safety Research

In the year and a half since FLI awarded the AI safety grants, our grantees have published over 40 academic publications and 10 news articles. Our grants funded 15 workshops and 5 academic courses, while the grantees participated in over 60 workshops and conferences.

These publications and workshops helped the AI safety movement continue to grow, and they garnered attention from both the academic and technology sectors. Some of the grantees' work, such as projects associated with the Future of Humanity Institute and by Joshua Greene also caught the attention of major news outlets including *The New York Times, Business Insider, Forbes, The Huffington Post, The Daily Mail, The Los Angeles Times, The Washington Post, Time,* and *GeekWire*.

In the past year, our grantees tackled some of the biggest questions in AI safety. For example, multiple teams looked at ensuring that machines act in accordance with human values. Stuart Russell and Owain Evans worked to develop improved methods of inverse

AI Safety Research, Cont.

reinforcement learning, Paul Christiano considered how we can supervise increasingly powerful AIs, and Francesca Rossi studied how to specify ethical laws through constraints.

Two of our grantees, Bas Steunebrink and Ramana Kumar, researched how machines can remain safe even when they have the power to modify their own code and encounter unforeseen situations.

FLI grant recipients participating in a panel discussion at AAAI.

And our grants aren't limited to technical research. Wendell Wallach brought together computer and social scientists to tackle the challenges of AI. Heather Roff looked at the concept of meaningful human control with regard to autonomous weapons, while Peter Asaro studied issues of liability and agency with autonomous systems. Moshe Vardi is organizing a multidisciplinary summit on job automation, and Nick Bostrom worked to derive policy desiderata for the transition into the machine intelligence era, focusing on efficiency, coordination, and the common good.

These examples are only a small subset of the research that is being done as a result of our grants program. This is the biggest grants program of its kind, and projects have already made an important impact in the AI community. We look forward to seeing the continued output from our grantees in the coming year, and we hope to expand our grants program to bring on more teams to help ensure AI is developed safely and beneficially for society.

Carnegie Mellon University
Carnegie Mellon University
Carnegie Mellon University

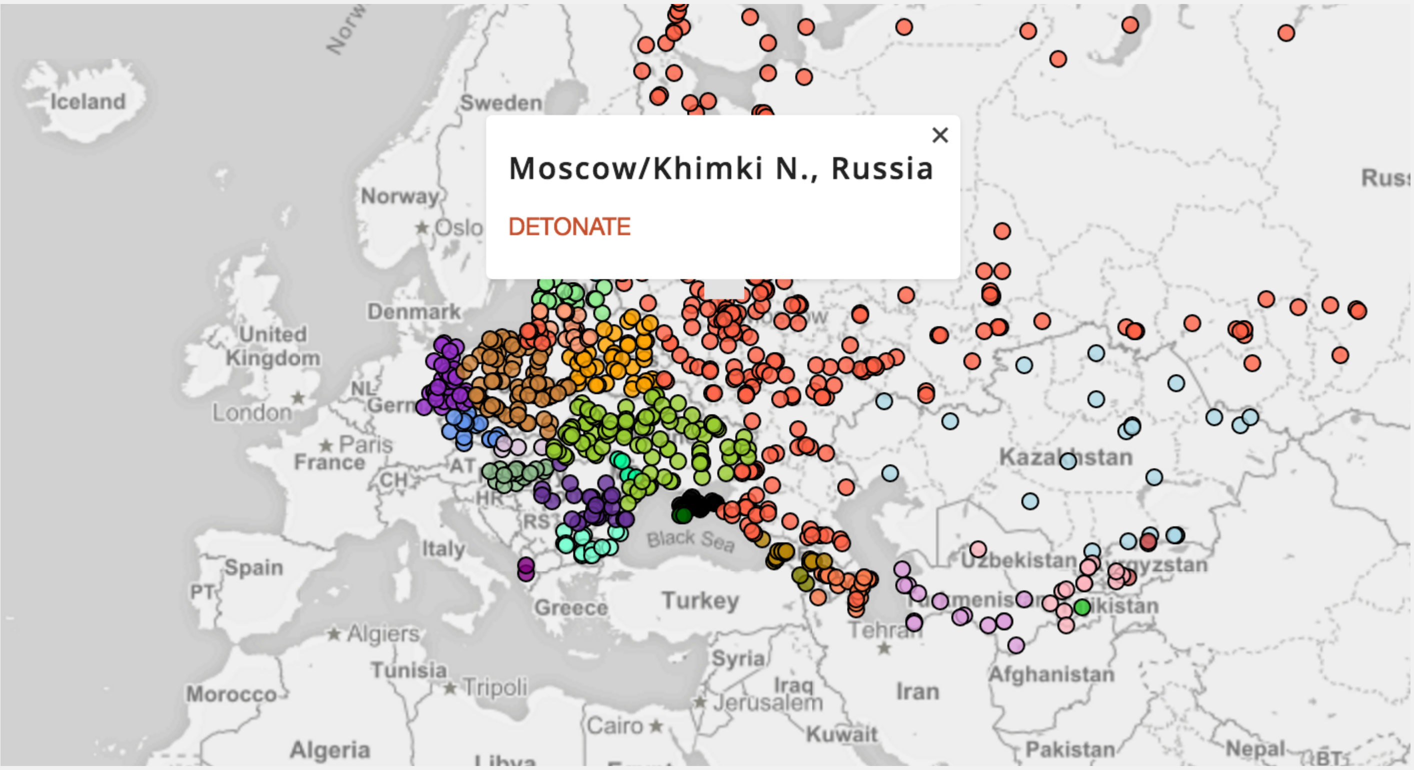
Online Outreach

2016 was a big year for FLI's online outreach initiatives. In just a year, the website and social media followers more than doubled, with the website attracting tens of thousands of visitors every month.

Our most effective web app was created in collaboration with Alex Wellerstein's NukeMap to show potential nuclear devastation based on the 1100 declassified U.S. nuclear targets from 1956. The app allows users to "detonate" various sized weapons at each target to see what the effect would be. Wellerstein also ran models of weather patterns over three days to see how nuclear fallout might travel and how many "friendly" countries would be at risk. The map was featured on *Motherboard, Gizmodo, Daily Mail, Popular Mechanics,* and *Fast Company*, and the page has received well over half a million visits to date.

We also developed a popular interactive timeline of nuclear close calls, and we created an app that highlights how many other great projects and programs could be funded with the money currently spent on nuclear weapons. Meanwhile, our page on the Benefits and Risks of Artificial Intelligence has grown increasingly popular every month, receiving hundreds of thousands of hits. We increased our news outreach through our own channels, but also via collaborations with *The Huffington Post, The World Post,* and *Futurism*.

In the fall, we launched a monthly podcast, which has seen growing success and represents one of the many efforts we're excited to do more with in the coming year.



The Beneficial AI Movement Keeps Growing

Among our greatest accomplishments and one of our biggest projects of 2016 was organizing the Beneficial AI Conference which took place at the start of 2017. The conference brought together some of the top minds in AI fields and confirmed that AI safety is part of the mainstream AI community. About 160 people attended BAI, which was twice the size of the Puerto Rico conference, requiring months of preparation. Many months were dedicated to organizing the event and laying the groundwork for the resulting Asilomar AI Principles.

As part of the BAI effort and with input from AI safety researchers, Richard Mallah created a Landscape of AI Safety and Beneficence Research. It organizes hundreds of technical research topics into a conceptual hierarchy and includes two levels of explanation for each research thread, annotated cross-references between topics, and nearly five hundred references. The data generated a survey paper, a flat map, and an interactive visualization.

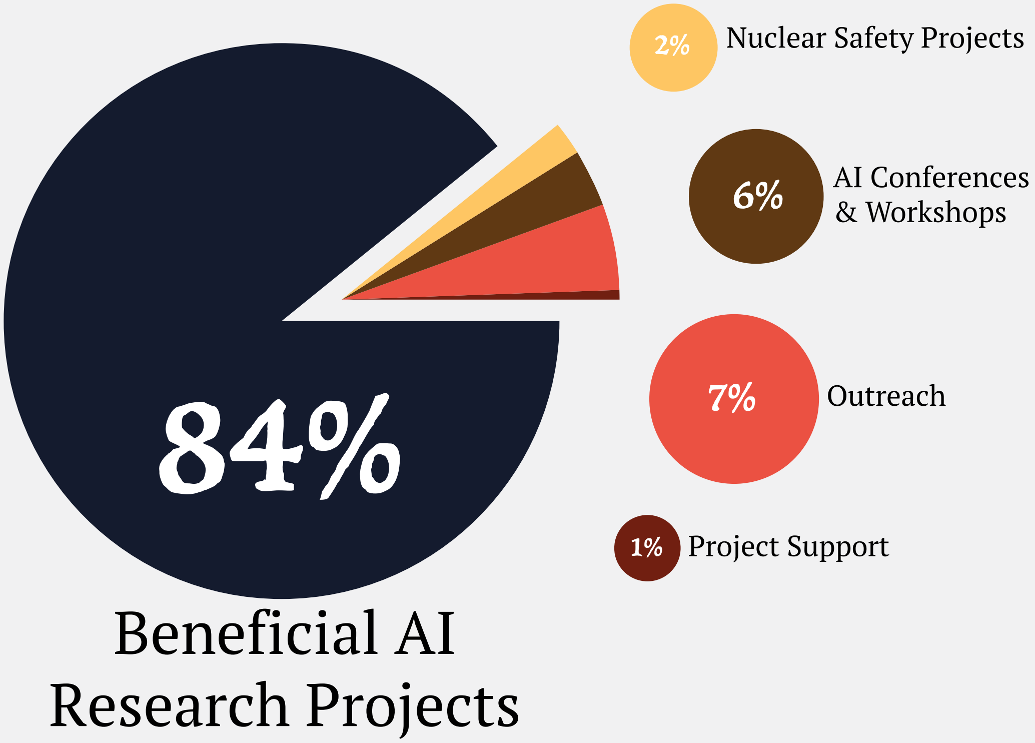
FLI members also participated in other AI conferences and workshops in 2016. Max Tegmark spoke at the NYU Symposium, The Future of AI, and he gave a joint talk with Meia Chita-Tegmark at another NYU conference, The Ethics of AI. FLI co-organized and sponsored the AI safety workshops Reliable Machine Learning in the Wild at ICML and Interpretable Machine Learning for Complex Systems at NIPS, and played an important role in the AI, Ethics, and Safety Workshop at the 30th annual AAAI conference. Viktoriya Krakovna co-ran an AI safety session at OpenAI's unconference, and she spoke at the ASU Governance of Emerging Technologies conference.

At the Effective Altruism Global (EAG) conference, Jaan Tallinn spoke on Philanthropy and Technology, Max Tegmark discussed the importance of dispelling myths about AI, and Richard Mallah gave a talk on mapping the landscape of AI safety research threads. In addition, Viktoriya Krakovna, Lucas Perry and Richard Mallah participated in EAGx Boston and Viktoriya discussed long-term AI risks on a panel at EAGx Oxford.

Volunteer Highlight: David Stanley

David is a postdoctoral researcher at Boston University specializing in computational neuroscience. This year, David oversaw numerous volunteers who clocked over 300 hours of volunteer work. He helped get nine of our most popular pages translated into eight languages, he's been working to expand our website and social media outreach to many other countries, and he played an integral role in helping to get our Trillion Dollar Nukes app developed. We look forward to working with him even more in 2017.

In 2016, we spent $2.6M, and most of it went to research...



We at FLI are proud that almost all of our income goes straight to our programs: in 2016, we spent $0 on fundraising and very little on administration. This efficiency is possible because our leadership and administration is mainly done by our unpaid board of directors and volunteers, and we have only a very small group of full-time employees overseeing outreach and other projects.

Most of the 2.6 million dollars spent in 2016 went to the 37 research teams around the world to whom we awarded grants for keeping AI beneficial: most of these payments are first installments to multi-year projects, which will continue to receive funding for 1-2 more years. Our outreach spending (7%) was mainly on our website, futureoflife.org, and our conference spending (6%) was mainly preparing for the 2017 Asilomar meeting.

2016 AI Safety Publications

Achim, T., Sabharwal, A., and Ermon, S. *Beyond parity constraints: Fourier analysis of hash functions for inference.* Proceedings of The 33rd International Conference on Machine Learning, pages 2254–2262, 2016. (Grant Recipient: Stefano Ermon)

This team presents progress on a novel approach to inference that's conducive both to formal guarantees of accuracy and to verification.

Armstrong, Stuart and Orseau, Laurent. *Safely Interruptible Agents*. Uncertainty in Artificial Intelligence (UAI) 2016. https://www.fhi.ox.ac.uk/interruptibility/. (Grant Recipient: Nick Bostrom)

This paper describes a new method for getting reinforcement learners to ignore system interruptions in their training data so they can be agnostic about system interruptions in their decisions. This research provides a method of neutralizing certain instrumental incentives that might be applicable in some form to other instrumental incentives in the future. The coauthorship by Google DeepMind, and the subsequent citations by various news outlets, including *Business Insider* and *Forbes*, has helped to mainstream concerns around scalable control.

Asaro, P. *The Liability Problem for Autonomous Artificial Agents*, Proceedings of the AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University, Stanford, CA, March 21-23, 2016. https://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12699 (Grant Recipient: Peter Asaro)

Laying a foundation for exploring the fundamental concepts of autonomy, agency, and liability as regards autonomous agents, this paper aims to helps prevent bad effects from falling through the cracks.

Bai, Aijun and Russell, Stuart. *Markovian State and Action Abstractions in Monte Carlo Tree Search*. In Proc. IJCAI16, New York, 2016. https://people.eecs.berkeley.edu/~russell/papers/ijcai16-markov.pdf (Grant Recipient: Stuart Russell)

This team introduces a significant optimization, summarizing very similar states together, allowing artificial agents to have further foresight when considering their next actions.

Boddington, Paula. *EPSRC Principles of Robotics: Commentary on safety, robots as products, and responsibility.* Ethical Principles of Robotics, special issue, 2016. http://www.tandfonline.com/eprint/7ifqJVESVzwkFzxVaGwS/full (Grant Recipient: Michael Wooldridge)

This paper analyzes a UK council's robotics principles recommendations with respect to definitions, gaps, contextualization, and lifecycle to help ensure such guidelines promote beneficence in realistic scenarios.

Boddington, Paula. *The Distinctiveness of AI Ethics, and Implications for Ethical Codes*. Presented at IJCAI-16 Workshop 6 Ethics for Artificial Intelligence, New York, July 2016. (Grant Recipient: Michael Wooldridge)

The author analyzes contextual considerations particular to creating guidelines and codes of conduct for the field, to help ensure that such guidelines promote beneficence in realistic scenarios.

Bostrom, N. *Strategic Implications of Openness in AI Development,* Technical Report #20161, Future of Humanity Institute, Oxford University: pp. 126, 2016. https://www.fhi.ox.ac.uk/wp-content/uploads/openness.pdf (Grant Recipient: Nick Bostrom)

This provides analysis of the strategic safety and security aspects of different kinds of sharing and openness in AI development, providing guidance on strategic coordination issues relevant to safety. Mapping out these non-obvious game theoretic dynamics should help inform the difficult coordination issues within each of the growing list of AI organizations concerned about safety and beneficence. This research also provides a framework for those interested in coordinating beneficial AGI in a global manner.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. *Robust Covariate Shift Regression*. International Conference on Artificial Intelligence and Statistics (AISTATS), 2016. http://jmlr.org/proceedings/papers/v51/chen16d.pdf (Grant Recipient: Brian Ziebart)

This effort gets regression to adapt to normally-problematic sample selection biases, aiming to improve the robustness of agents that use regression.

Conitzer, V., Sinnott-Armstrong, W., Borg, J.S., Deng, Y., and Kramer, M. *Moral Decision Making Frameworks for Artificial Intelligence*. (Preliminary version.) To appear in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) Senior Member / Blue Sky Track, San Francisco, CA, USA, 2017. https://users.cs.duke.edu/~conitzer/moralAAAI17.pdf (Grant Recipient: Vincent Conitzer)

This team explores encouraging mutual trust in a game-theoretic manner, as well as the machine learning of moral judgment, seeking footings for grounded computational ethics.

Critch, Andrew. *Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents*. 2016. http://intelligence.org/files/ParametricBoundedLobsTheorem.pdf (Grant Recipient: Benya Fallenstein)

This paper develops the field of open source game theory as it applies to limited agents, facilitating future explorations of making an agent that creates another agent it can have confidence in.

Evans, O., Stuhlmuller, A., and Goodman, N.D. *Learning the Preferences of Ignorant, Inconsistent Agents*. 2015. http://arxiv.org/pdf/1512.05832.pdf (Grant Recipient: Owain Evans)

This team introduces a framework for accounting for, and attempting to correct for, cognitive limitations in the agents from which it is to learn goals by watching their behavior. This can help enable machines to better understand what the operators really want. This framework has already led to accounting for additional human limitations when performing metapreference resolution. Accounting for humans' limitations when figuring out what they want deep down, and doing so in a framework that learns from their behaviors and artifacts, is an important step toward the now-classic vision of coherent extrapolated volition.

Evans, O., Stuhlmuller, A., and Goodman, N.D. *Learning the Preferences of Bounded Agents*. NIPS Workshop on Bounded Optimality, 2015. http://stuhlmueller.org/papers/preferencesnipsworkshop2015.pdf (Grant Recipient: Owain Evans)

In this paper, additional cognitive biases are accounted for in the agents from which to learn goals, helping machines better understand what the operators are aiming for.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. *Adversarial Multiclass Classification: A Risk Minimization Perspective*. Neural Information Processing Systems (NIPS), 2016. https://papers.nips.cc/paper/6088-adversarial-multiclass-classification-a-risk-minimization-perspective.pdf (Grant Recipient: Brian Ziebart)

This team reformulates adversarial classification algorithms into forms conducive to a number of consistency guarantees, potentially enabling a wider range of verification techniques.

Fulton, Nathan and Platzer, André. *A logic of proofs for differential dynamic logic: Toward independently checkable proof certificates for dynamic logics*. Jeremy Avigad and Adam Chlipala, editors, Proceedings of the 2016 Conference on Certified Programs and Proofs, CPP 2016, St. Petersburg, FL, USA, January 18-19, 2016, pp. 110-121. ACM, 2016. http://nfulton.org/papers/lpdl.pdf (Grant Recipient: Andre Platzer)

This creates a cleaner and more extensible way to model and verify implementation correctness of physical-computational systems, by establishing a more robust core infrastructure.

Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N., and Taylor, J. *Asymptotically Coherent, Well Calibrated, Self-trusting Logical Induction*. Working Paper (Berkeley, CA: Machine Intelligence Research Institute). 2016. https://arxiv.org/pdf/1609.03543.pdf (Grant Recipient: Benya Fallenstein)

In this paper, a robust new algorithm for inferring the likelihood of logical statements is developed, which among other applications, can facilitate metareasoning about potential successors and introduce fault-tolerance during long-running calculations. Accounting for logical induction can revolutionize agent-modeling fields like economics and game theory, as well as the safety work based on those. Most crucially, logical induction is expected to facilitate metareasoning about potential successors.

Garrabrant, S., Fallenstein, B., Demski, A., and Soares, N. *Inductive Coherence*. arXiv:1604.05288 [cs.AI]. 2016. https://arxiv.org/pdf/1604.05288v3.pdf (Grant Recipient: Benya Fallenstein)

This paper provides appropriate constraints on finite approximations to probability distributions over logical statements, helping establish some groundwork for metareasoning around long-running deductions.

Garrabrant, S., Soares, N., and Taylor, J. *Asymptotic Convergence in Online Learning with Unbounded Delays*. arXiv:1604.05280 [cs.LG]. 2016. https://arxiv.org/pdf/1604.05280.pdf (Grant Recipient: Benya Fallenstein)

This team analyzes how quickly algorithms that forecast the results of computations can achieve good behavior, helping establish some groundwork useful for a logical induction algorithm.

Greene, J. D. *Our driverless dilemma*. Science, 352(6293), 1514-1515. 2016. (Grant Recipient: Francesca Rossi)

Analyzing interactions between utilitarianism and game theory in determining computational ethics, the author helps tangibly bridge validation and beneficence in the real-world.

Greene, J., Rossi, F., Venerable, K.B., Tasioulas, J., and Williams, B. *Embedding Ethical Principles in Collective Decision Support Systems*. Thirtieth AAAI Conference on Artificial Intelligence. March 2016. www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12457/12204 (Grant Recipient: Francesca Rossi)

This paper investigates considerations for negotiating acceptably ethical outcomes among systems representing multiple parties, a potentially equitable computational ethics mechanism.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. *Cooperative Inverse Reinforcement Learning*. Neural Information Processing Systems (NIPS), 2016. https://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf (Grant Recipient: Stuart Russell)

This team introduces a significant new algorithm for machines and humans to collaborate on uncovering the human's value function, helping enable machines to better understand what their operators really want. This adds a key method and modality for training assistive agents, uncovering preferences, objectives, and values. It provides a solid framework that can be expanded with new capabilities for learning both instrumental and deeper objectives.

Hsu, L.K., Achim, T., and Ermon, S. *Tight variational bounds via random projections and i-projections*. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pages 1087–1095, 2016. https://arxiv.org/pdf/1510.01308v1.pdf (Grant Recipient: Stefano Ermon)

Using a new coarse-grain discrete summarization of a statistical distribution, with provable guarantees on its approximation quality, this team opens the door to such models becoming conducive to many types of proofs and verification. Being able to take complex, fine-grained, or continuous models, as are common in statistical machine learning, and reframe them as discrete summarizations can help getting past one of the major hurdles in end-to-end verification of machine learning systems.

Khani, F., Rinard, M., and Liang, P. *Unanimous prediction for 100% precision with application to learning semantic mappings*. Association for Computational Linguistics (ACL), 2016. http://arxiv.org/pdf/1606.06368v2.pdf (Grant Recipient: Percy Liang)

This paper introduces a method for establishing when an ensemble must be correct, otherwise communicating uncertainty, providing robustness against false positives and potentially enabling proofs useful to a wider range of verification techniques.

Kim, C., Sabharwal, A., and Ermon, S. *Exact sampling with integer linear programs and random perturbations*. Proc. 30th AAAI Conference on Artificial Intelligence, 2016. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12105/12087 (Grant Recipient: Stefano Ermon)

This team develops a novel way to sample probabilistic graphical machine learning models more scalably, thereby also helping scale our understanding of bounds, behavior, and approximate guarantees to more complex models closer to runtime.

Leike, J., Taylor, J., and Fallenstein, B. *A Formal Solution to the Grain of Truth Problem*. Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), edited by Alexander Ihler and Dominik Janzing, 427–436. Jersey City, New Jersey, USA. 2016. http://www.auai.org/uai2016/proceedings/papers/87.pdf (Grant Recipient: Benya Fallenstein)

This team introduces a technique for finding criteria for correctly identifying the multiplayer game an agent is in and for achieving a mutually beneficial equilibrium in that game, building on a recent foundation of game theory to aid decision theory, bounded rationality, and engineering of cooperative behavior.

Liu, C., Hamrick, J. B., Fisac, J. F., Dragan, A. D., Hendrick, J. K., Sastry, S. S., and Griffiths, T. L. *Goal inference improves objective and perceived performance in human robot collaboration*. In Proc. AAMAS16, Singapore, 2016. http://www.jesshamrick.com/publications/pdf/Liu2016-Goal_Inference_Improves_Objective.pdf (Grant Recipient: Stuart Russell)

This paper investigates a human-machine collaboration scheme with goal anticipation and dynamic replanning, which lets machines interact with humans more smoothly while they learn what humans want.

Perera, V., Selveraj, S.P., Rosenthal, S., and Veloso, M. *Dynamic Generation and Refinement of Robot Verbalization*. Proceedings of RO-MAN'16, the IEEE International Symposium on Robot and Human Interactive Communication, Columbia University, NY, August, 2016. http://www.cs.cmu.edu/~mmv/papers/16roman-verbalization.pdf (Grant Recipient: Manuela Veloso)

This explores a step to generalizing verbal interaction between embodied robots and humans, fostering models of interpretability so a broader range of people understand system intent.

Pistono, F and Yampolskiy, RV. *Unethical research: How to create a malevolent artificial intelligence*. 25th International Joint Conference on Artificial Intelligence (IJCAI-16), Ethics for Artificial Intelligence Workshop (AI-Ethics-2016). https://arxiv.org/ftp/arxiv/papers/1605/1605.02817.pdf (Grant Recipient: Seth Baum)

Analyzing scenarios where designers purposely instill malevolence into an agent, this paper helps identify some potential intervention points where purposely dangerous agents may be detected or thwarted.

Rosenthal, S., Selvarej, S.P., and Veloso, M. *Verbalization: Narration of Autonomous Mobile Robot Experience*, In Proceedings of IJCAI'16, the 26th International Joint Conference on Artificial Intelligence, New York City, NY, July, 2016. http://www.cs.cmu.edu/ mmv/papers/16ijcai-verbalization.pdf (Grant Recipient: Manuela Veloso)

This team introduces the concept of verbalization space for representing modifier parameters for generation of agent explanations, fostering models of interpretability so a broader range of people understand system intent in a broader range of contexts.

Rossi, F. *Ethical Preference-Based Decision Support System*. Proc. CONCUR 2016, Springer. 2016. http://drops.dagstuhl.de/opus/volltexte/2016/6187/pdf/LIPIcs-CONCUR-2016-2.pdf (Grant Recipient: Francesca Rossi)

The author investigates how preferences and meta-preferences can be used for representing morality and for collective ethical decision-making, exploring some prerequisites and mechanisms for practical aggregate value alignment.

Rossi, F. *Moral preferences*, Proc. IJCAI 2016 workshop on AI and ethics, and Proc. IJCAI 2016 workshop on multidisciplinary approaches to preferences. 2016. https://intelligence.org/files/csrbai/pref-eth1.pdf (Grant Recipient: Francesca Rossi)

In this paper, adaptation of preference frameworks and meta-preferences to modeling ethical theories is investigated, to explore prerequisites and mechanisms for practical value specification.

Siddiqui, A., Fern, A., Dietterich, T. G., and Das, S. *Finite Sample Complexity of Rare Pattern Anomaly Detection*. Proceedings of UAI-2016 (pp. 10). 2016. (Grant Recipient: Thomas Dietterich)

Helping to establish more rigorous understanding of the process of recognizing when things are unusual, this team aides AIs' increasing contextual awareness in a way that may lead to formal proofs or even verifiability.

Steinhardt, J. and Liang, P. *Unsupervised Risk Estimation with only Conditional Independence Structure*. Neural Information Processing Systems (NIPS), 2016. https://papers.nips.cc/paper/6201-unsupervised-risk-estimation-using-only-conditional-independence-structure (Grant Recipient: Percy Liang)

This paper introduces a method for estimating a model's test error from unlabeled data, providing another way of identifying uncertainty after deployment.

Steunebrink, B.R., Thorisson, K., and Schmidhuber, J. *Growing Recursive Self-Improvers*. Proceedings of the 9th Conference on Artificial General Intelligence (AGI 2016), LNAI 9782, pages 129-139. Springer, Heidelberg. 2016. http://people.idsia.ch/~steunebrink/Publications/AGI16_growing_recursive_self-improvers.pdf (Grant Recipient: Bas Steunebrink)

This team investigates experience-based ideation and testing of self-modifications, aiming to provide a more empirical approach to safer recursive self-improvement. This approach enables the exploration of counterfactuals that purely analytic solutions to proposed self-improvements could not address. Moreover, it provides a relatively easy framework for incrementally testing the safety of self-modifications.

Taylor, Erin. *The Threat-Response Model: Ethical Decision in the Real World.*

This paper investigates how a quantitative model used in security-related decision making can be extended to broader ethical decisions, enabling exploration of a possible framework for integrating planning and value alignment.

Taylor, Jessica. *Quantilizers: A Safer Alternative to Maximizers for Limited Optimization*. 2nd International Workshop on AI, Ethics and Society at AAAI-2016. Phoenix, AZ. 2016. www.aaai.org/ocs/index.php/WS/AAAIW16/paper/download/12613/12354 (Grant Recipient: Benya Fallenstein)

This paper proposes a method for satisficing an objective rather than over-optimizing for it, aiming to support environmental low-impact of powerful advanced agents.

Thorisson, K.R., Bieger, J., Thorarensen, T., Sigurðardóttir, J.S., and Steunebrink, B.R. *Why Artificial Intelligence Needs a Task Theory (And What It Might Look Like)*. Proceedings of the 9th Conference on Artificial General Intelligence (AGI 2016), LNAI 9782, pages 118-128. Springer, Heidelberg. 2016. http://people.idsia.ch/~steunebrink/Publications/AGI16_task_theory.pdf (Grant Recipient: Bas Steunebrink)

This team discusses the motivation and desiderata for a way to model tasks, a theory that would aide in formalizing usage of non-optimizing satisfaction, and would help in standardizing measures of capability across different AGI systems.

Thorisson, K.R., Kremelberg, D., Steunebrink, B.R., and Nivel, E. *About Understanding*. Proceedings of the 9th Conference on Artificial General Intelligence (AGI 2016), LNAI 9782, pages 106-117. Springer, Heidelberg. 2016. http://link.springer.com/chapter/10.1007/978-3-319-41649-6_11 (Grant Recipient: Bas Steunebrink)

Investigating desiderata and mechanisms for rich pragmatic understanding through multimodal interaction, as this team does, supports common-sense reasoning and contextual awareness within advanced agents.

Tossou, A.C.Y. and Dimitrakakis, C. *Algorithms for Differentially Private Multi-Armed Bandits*. Proc. 13th AAAI Conf. on Artificial Intelligence (AAAI 2016), 2016. www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11906/11846 (Grant Recipient: David Parkes)

This paper introduces algorithms for designing privacy-respecting experiments where individual records are connected to private data, aiding security of machine learning and cooperation between a broader set of agents.

Wellman, MP and Rajan, U. *Ethical issues for autonomous trading agents*. IJCAI-16 Workshop on Ethics for Artificial Intelligence, July 2016. (Grant Recipient: Michael Wellman)

This duo explores the various ways financial agents can unethically game market and societal systems, providing scenario analysis for risks of autonomous agents and suggestions for where new rules are needed.

Yampolskiy, RV. *Taxonomy of pathways to dangerous AI*. 30th AAAI Conference on Artificial Intelligence (AAAI-2016), 2nd International Workshop on AI, Ethics and Society (AI Ethics Society 2016). www.aaai.org/ocs/index.php/WS/AAAIW16/paper/download/12566/12356 (Grant Recipient: Seth Baum)

This paper lays out a variety of scenarios that lead to dangerous AI, potentially informing those considering fostering AGI as to coordinative and technical risks they can work to mitigate.

Zhang, Rubinstein, B.I.P., and Dimitrakakis, C. *On the Differential Privacy of Bayesian Inference*. Proc. 13th AAAI Conf. on Artificial Intelligence (AAAI 2016), 2016. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/12129/11885 (Grant Recipient: David Parkes)

This introduces methods to communicate conditional probabilistic inference while maintaining privacy of underlying data, aiding security of machine learning and cooperation between a broader set of agents.

Zhao, S., Chaturapruek, S., Sabharwal, A., and Ermon, S. *Closing the gap between short and long xors for model counting*. Thirtieth AAAI Conference on Artificial Intelligence, 2016. https://ai2-website.s3.amazonaws.com/publications/shortXors-AAAI2016.pdf (Grant Recipient: Stefano Ermon)

This paper analyzes how approximate model counting can be subject to more specific bounds, which can more efficiently make probabilistic inference robust.

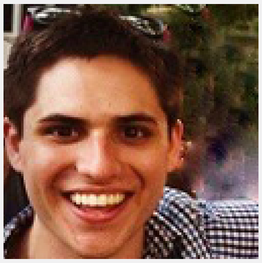
Thank You!

We would like to extend a special thank you to the donors who made all these great accomplishments possible: 100% of the $2.6M described on the previous page came from philanthropic donations. We are especially grateful to Elon Musk for his generous donation, which enabled the continuation of our beneficial AI grants program.

We are also deeply thankful to the Open Philanthropy Project, Jaan Tallinn, Matt Wage, Alexander Tamas, Brian Corcoran, Jacob Trefethen and many other donors whose generous support helped make possible everything we have done so far.

Open Philanthropy Project

Jaan Tallinn



Matt Wage

Alexander Tamas

Jacob Trefethen

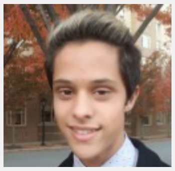
An additional thank you must also go out to our smaller donors and our volunteers who made FLI a priority this year. Without your help, we could never have accomplished as much as we did.

Core Team & Volunteers

Ariel Conn

Tucker Davey

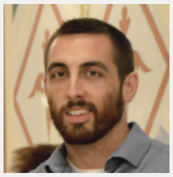
Richard Mallah


Lucas Perry

David Stanley

Eric Gastfriend

Grzegorz Orwinski


Jacob Beebe

Na Li (Lina)

Rafael Martinez-Galarza

Vera Koroleva

Zara Yaqoob

future of life INSTITUTE

This is What Nuclear Fallout Could Look Like

MOTHERBOARD

'Press the big red button': Computer experts want kill switch to stop robots from going rogue

The Washington Post

Future of Life Institute, Contributor

Research and initiatives for safeguarding life and developing optimistic visions of the future.

THEWORLDPOST

A PARTNERSHIP OF THE HUFFINGTON POST AND BERGGRUEN INSTITUTE

Artificial Intelligence And The King Midas Problem

Autonomous Weapons Are Already Here

POPULAR SCIENCE

Simulate the End of the World With This Interactive Map of U.S. Nuclear Targets

POPULAR MECHANICS

The New York Times

Should Your Driverless Car Hit a Pedestrian to Save Your Life?

Vox

This 3-minute cartoon explains why nuclear weapons still pose a very real threat

Making a Difference