Before we begin...

Wi-Fi Connection
Username: Marriott Bonvoy (no password needed)

Conference Recording
The recording will be available after the event on our official YouTube channel: ***AIDXTECH***.

Sustainable Name Tag
Your name tag is printed on seed paper, you can plant it and watch it grow!

1. Soak the paper for 30 minutes, then tear into pieces.

2. Plant, water regularly, and keep in a cool, shaded spot.

3. Sprouts will appear in 7-15 days.

Photography & Media Notice

Please be advised that photographs and video footage will be taken during this event for use in publications, digital platforms, and other publicity materials by the organizers. By attending this event, you consent to the use of your image and likeness in such media. This is in accordance with the Personal Data Protection Act (PDPA). If you have any concerns or wish to opt out, please approach the registration desk or a member of the event staff.

Large Model Safety Workshop 2025

AIDX TECH PTE. LTD.

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Artificial Intelligence International Institute

CyberSG R&D Programme Office

安远 AI
CONCORDIA AI

SINGAPORE EXHIBITION & CONVENTION BUREAU

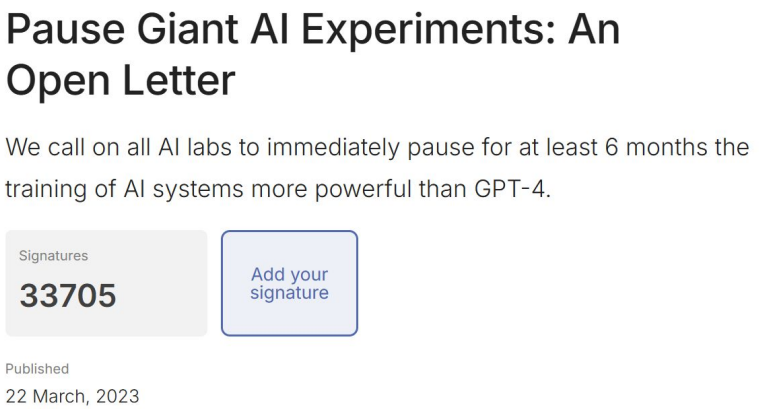
A Story

In this classic sci-fi novel (“The Day of the Triffids”) , the world is warned about the potential danger of Triffids — tall, carnivorous, mobile plants with venomous stingers. Scientists study them, and people recognize them as a potential threat, but because they are largely controlled and cultivated for their valuable oil, society stops worrying.

However, a sudden catastrophe … and civilization collapses as the forgotten danger takes over.

The Worry Over AI

March, 2023



2025

January: DeepSeek-R1, Qwen2.5-Max

February: Grok-3, Claude 3.7

March: Manus, QwQ-32B

April: Llama 4 Series, GPT-4.5, GPT-o3-mini

The Worry Over AI

ChatGPT

Thanks. I need a slightly different one, where the people was warned about the danger, worried first, but then forgot about it and eventually paid for it.

A great example that fits your request is *The Day of the Triffids* by John Wyndham.

Summary:

In this classic sci-fi novel, the world is warned about the potential danger of Triffids—tall, carnivorous, mobile plants with venomous stingers. Scientists study them, and people recognize them as a potential threat, but because they are largely controlled and cultivated for their valuable oil, society stops worrying. Over time, the Triffids are treated as just another part of life, their danger forgotten.

However, a sudden catastrophe—a mysterious green meteor shower—renders most of humanity blind. With

Program

Morning

8:40-9:20: Dawn Song, UC Berkeley
9:20-10:00: Gagandeep Singh, UIUC
10:00-10:40: Xin Eric Wang, UC Santa Cruz

10:40-11:00: Coffee Break

11:00-11:40: Christopher Manning, Stanford
11:40-12:20: Yang Zhang from CISPA
12:20-13:00: Mohit Bansal, University of North Carolina at Chapel Hill

Afternoon

13:00-14:00: Lunch

14:00-14:40: Yoshua Bengio, U. of Montreal
14:40-15:20: Bo Li from UIUC
15:20-16:00: Maksym Andriushchenko, EPFL

16:00-16:30: Coffee Break

16:30-17:15: Panel Discussion

Dawn Song is a Professor in Computer Science at UC Berkeley and Co-Director of Berkeley Center for Responsible Decentralized Intelligence. Her research interest lies in AI and deep learning, security and privacy, and decentralization technology. She is the recipient of various awards including the MacArthur Fellowship, the Guggenheim Fellowship, the NSF CAREER Award, the Alfred P. Sloan Research Fellowship, the MIT Technology Review TR-35 Award, ACM SIGSAC Outstanding Innovation Award, and more than 10 Test-of-Time Awards and Best Paper Awards from top conferences in Computer Security and Deep Learning. She has been recognized as Most Influential Scholar (AMiner Award), for being the most cited scholar in computer security. She is an ACM Fellow and an IEEE Fellow. She obtained her Ph.D. degree from UC Berkeley. She is also a serial entrepreneur and has been named on the Female Founder 100 List by Inc. and Wired25 List of Innovators.

Gagandeep Singh is an Assistant Professor in the Siebel School of Computing and Data Science at the University of Illinois Urbana-Champaign (UIUC). He is also part of the Science and Technology group of the Institute of Government and Public Affairs at the University of Illinois. His research combines ideas from formal logic, machine learning, and systems research to construct intelligent systems with formal guarantees about their behavior and safety. His group at UIUC has received several awards and fellowships, including the NSF Career, Google Research Scholar, Amazon Research, and the Qualcomm Innovation Fellowship. He has served on the program committee for top conferences in machine learning, security software engineering, and programming languages, such as NeurIPS, ICML, ICLR, CVPR, ICCV, CCS, ICSE, ASPLOS, CAV, POPL, and PLDI.

Xin (Eric) Wang is an Assistant Professor of Computer Science and Engineering at UC Santa Cruz and the Head of Research at Simular. His research interests include Natural Language Processing, Computer Vision, and Machine Learning, with an emphasis on Multimodal Reasoning and AI Agents. He worked at Google Research, Facebook AI Research (FAIR), Microsoft Research, and Adobe Research. Eric has served as Area Chair for conferences such as ACL, NAACL, EMNLP, ICLR, and NeurIPS, as well as a Senior Program Committee for AAAI and IJCAI. He organized workshops and tutorials at conferences such as ACL, NAACL, CVPR, and ICCV. He has received several awards and recognitions for his work, including CVPR Best Student Paper Award, Google Research Faculty Award, Amazon Alexa Prize Awards, Cisco Research Award, eBay Research Awards, and various faculty research awards from Adobe, Snap, Microsoft, Cybever, etc.

Christopher Manning is the inaugural Thomas M. Siebel Professor in Machine Learning in the Departments of Linguistics and Computer Science at Stanford University, a Founder and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and was Director of the Stanford Artificial Intelligence Laboratory (SAIL) from 2018–2025. From 2010, Manning pioneered Natural Language Understanding and Inference using Deep Learning, with impactful research on sentiment analysis, paraphrase detection, the GloVe model of word vectors, attention, neural machine translation, question answering, self-supervised model pre-training, tree-recursive neural networks, machine reasoning, summarization, and dependency parsing, work for which he has received two ACL Test of Time Awards and the IEEE John von Neumann Medal (2024). He earlier led the development of empirical, probabilistic approaches to NLP, computational linguistics, and language understanding, defining and building theories and systems for natural language inference, syntactic parsing, machine translation, and multilingual language processing, work for which he won ACL, Coling, EMNLP, and CHI Best Paper Awards. In NLP education, Manning coauthored foundational textbooks on statistical NLP (Manning and Schütze 1999) and information retrieval (Manning, Raghavan, and Schütze, 2008), and his online CS224N Natural Language Processing with Deep Learning course videos have been watched by hundreds of thousands. In linguistics, Manning is a principal developer of Stanford Dependencies and Universal Dependencies, and has authored monographs on ergativity and complex predicates. He is the founder of the Stanford NLP group (@stanfordnlp) and was an early proponent of open source software in NLP with Stanford CoreNLP and Stanza. He is an ACM Fellow, an AAAI Fellow, and an ACL Fellow, and was President of the ACL in 2015. Manning earned a B.A. (Hons) from The Australian National University, a Ph.D. from Stanford in 1994, and an Honorary Doctorate from U. Amsterdam in 2023.

Yang Zhang is a tenured faculty member at CISPA Helmholtz Center for Information Security, Germany. His research concentrates on trustworthy machine learning including privacy, security and more recently LLM safety. Moreover, he works on measuring and understanding misinformation and unsafe content like hateful memes on the Internet. His research has been featured in major media outlets including the Washington Post and New Scientist. He has received the NDSS 2019 distinguished paper award and the CCS 2022 best paper award runner-up.

Dr. Mohit Bansal is the John R. & Louise S. Parker Distinguished Professor and the Director of the MURGe-Lab (UNC-NLP Group) in the Computer Science department at UNC Chapel Hill. He received his PhD from UC Berkeley in 2013 and his BTech from IIT Kanpur in 2008. His research expertise is in natural language processing and multimodal machine learning, with a particular focus on multimodal generative models, grounded and embodied semantics, reasoning and planning agents, faithful language generation, and interpretable, efficient, and generalizable deep learning. He is a AAAI Fellow and recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE), IIT Kanpur Young Alumnus Award, DARPA Director's Fellowship, NSF CAREER Award, Google Focused Research Award, Microsoft Investigator Fellowship, Army Young Investigator Award (YIP), DARPA Young Faculty Award (YFA), and outstanding paper awards at ACL, CVPR, EACL, COLING, CoNLL, and TMLR. He has been a keynote speaker for the AACL 2023, CoNLL 2023, and INLG 2022 conferences. His service includes EMNLP and CoNLL Program Co-Chair, and ACL Executive Committee, ACM Doctoral Dissertation Award Committee, ACL Americas Sponsorship Co-Chair, and Associate/Action Editor for TACL, CL, IEEE/ACM TASLP, and CSL journals.

Yoshua Bengio is Full Professor of Computer Science at Université de Montreal, as well as the Founder and Scientific Director of Mila and a Canada CIFAR AI Chair. Considered one of the world’s leaders in Artificial Intelligence and Deep Learning, he is the recipient of the 2018 A.M. Turing Award, considered like the "Nobel prize of computing". He is also the most cited computer scientist worldwide.

Professor Bengio is a Fellow of both the Royal Society of London and Canada, an Officer of the Order of Canada, a Knight of the Legion of Honor of France, a member of the UN’s Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology and chairs the International AI Safety Report.

Dr. Bo Li is an Associate Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. She is the recipient of the IJCAI Computers and Thought Award, Alfred P. Sloan Research Fellowship, IEEE AI's 10 to Watch, NSF CAREER Award, MIT Technology Review TR-35 Award, Dean's Award for Excellence in Research, C.W. Gear Outstanding Faculty Award, Intel Rising Star Award, Symantec Research Labs Fellowship, Rising Star Award, Research Awards from Tech companies such as Amazon, Meta, Google, Intel, IBM, and eBay, JPMC, Oracle, and best paper awards at several top machine learning and security conferences. Her research focuses on both theoretical and practical aspects of trustworthy machine learning, which is at the intersection of machine learning, security, privacy, and game theory. Her work has been featured by several major publications and media outlets, including Nature, Wired, Fortune, and New York Times.

Maksym Andriushchenko is a postdoctoral researcher at EPFL and an ELLIS Member. He has worked on AI safety with leading organizations in the field (OpenAI, Anthropic, UK AI Safety Institute, Center for AI Safety, Gray Swan AI). He obtained a PhD in machine learning from EPFL in 2024 advised by Prof. Nicolas Flammarion. His PhD thesis was awarded with the Patrick Denantes Memorial Prize for the best thesis in the CS department of EPFL and was supported by the Google and Open Phil AI PhD Fellowships.

Large Model Safety Workshop 2025

AIDX TECH PTE. LTD.

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Artificial Intelligence International Institute

CyberSG R&D Programme Office

安远 AI
CONCORDIA AI

SINGAPORE EXHIBITION & CONVENTION BUREAU