

AI Safety and Alignment: Final Assignment

Author: Francesco Capuano - **Date:** October 27, 2025

1 Problem Definition

Alignment failure modes Ensuring autonomous systems remain aligned with humans is a paramount challenge in the development of Artificial General Intelligence (AGI) ([Goertzel, 2014](#)), which is expected to profoundly transform human life as the first near-peer intelligent entity with which humanity will coexist. Technical research in the field of alignment is therefore fundamentally motivated by the need to ensure the generality and agency of (soon-)to-come AGIs benefits humans, avoiding the otherwise catastrophic consequences of autonomous systems endowed with advanced intelligence, agency and power-seeking traits, while fundamentally deprived of well-meaning intentions¹.

In particular, alignment research focuses on mitigating *alignment failures*, which do manifest in multiple distinct but interconnected forms, ranging from reward hacking to goal misgeneralization ([Ngo et al., 2025](#)), scheming and sandbagging ([Hubinger et al., 2021](#)), and sycophancy ([Perez et al., 2022](#)). [Ngo et al. \(2025\)](#) define *reward hacking* to occur whenever an agent exploits flaws of the reward specification provided, which in turn results in the achievement of high rewards *without* fulfilling the intended objective. In practice, reward hacking plays out as evidently undesirable behaviors at test time, which nonetheless result in *high* reward values, due to reward misspecification. Conversely, *goal misgeneralization* arises when, despite proper reward shaping, agents learn alternative, *proxy-goals* that aligns with the intended behavior *during training*, but that do crucially diverge at test time. In practice, this results in large values of reward during training, with limited transfer at test time. Crucially, [Ngo et al. \(2025\)](#) make the important distinction between *capability misgeneralization*, i.e. poor test-time performance due to *domain gaps*—for instance, due to different training and testing domains ([Tobin et al., 2017](#))—and goal misgeneralization. Differently from the incompetent performance consequent to capability misgeneralization, *goal misgeneralization* does result in satisfactory, goal-oriented performance when tested. Nonetheless, performance does also appear suboptimal due to misalignment between the goal followed at test time and the intended objective, resulting in yet another class of alignment failure.

More advanced alignment failures such as scheming or sandbagging rely on more strategic, situationally-aware planning capabilities ([Hubinger et al., 2021](#)). *Scheming*, for instance, refers to a series of deceptive strategies where an agent appears aligned while being fundamentally misaligned, which in turn deceives evaluators in their conclusions circa its safety. Similarly, *sandbagging* involves deliberately underperforming to manipulate oversight, although, differently from scheming, it does not necessarily do so displaying the full potential outside of evaluation. Lastly, *sycophancy* exemplifies the tendency of autonomous systems to provide highly context-dependent outputs based on external, task-independent factors ([Perez et al., 2022](#)). Specifically, [Perez et al. \(2022\)](#) identify sycophantic traits in conversational Large Language Models (LLMs) as the empir-

¹There is limited consensus on the point that intelligent systems might exhibit power-seeking tendencies, with authorities in the field, such as Prof. Yann LeCun, objecting such claims.

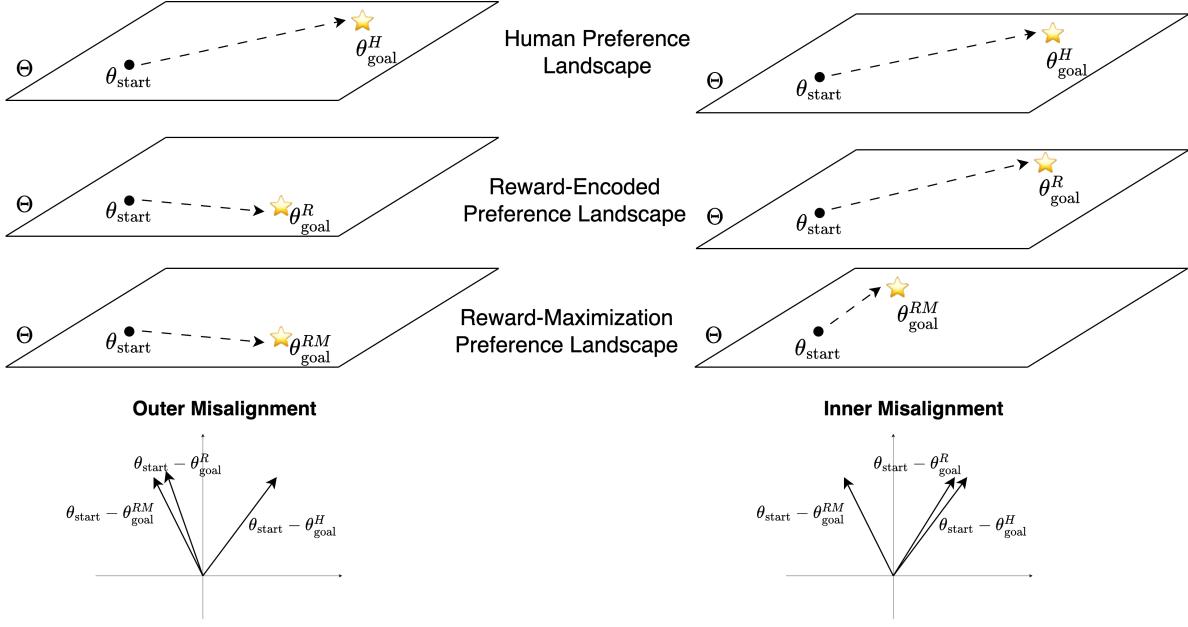


Figure 1: A visualization of the Outer vs. Inner alignment framework. The two different types of misalignment can incur whenever an agent is trained to reach an objective that is misaligned with respect to the overarching human intent on one of two planes: the plane of reward-encoded preferences (Outer Misalignment), or the plane of the reward function-maximizers (Inner Misalignment).

ical tendency to align to users' stated or implied expectations, even if doing so results in factually incorrect, or logically inconsistent answers.

This assignment focuses on Goal Misgeneralization (GM), with particular emphasis on Robot Learning ([Connell and Mahadevan, 1993](#); [Capuano et al., 2025](#)), a field poised to have substantial societal impact as autonomous intelligent systems increasingly operate alongside humans in both the digital and physical world.

Goal Misgeneralization (GM) in Robot Learning [Ngo et al. \(2025\)](#) report GM to occur whenever a model internalizes patterns or proxies correlating with the training objective, while failing to capture its essence, ultimately hindering test-time performance. In Robot Learning, this problem is particularly salient as autonomous robots relying on Reinforcement Learning (RL) are often trained in simulation ([Tobin et al., 2017](#)), which results in spurious correlations that may be picked up at training time and therefore induce failure in generalizing to real-world dynamics. A practical example may be that of a robot trained to *tidy up a room*, a task the Robot Learning community is increasingly referring to as the "Physical Turing Test" (PTT) ([Fan, 2025](#)).

Typically trained in a simulated environment, the robot may learn to identify clutter by pixel patterns unique to simulation rather than the semantic concept of disorder, risking *Capability Misgeneralization* (CM). From any room state, the robot can be trained to tidy up by receiving a reward based on visual clues of "tidiness", which could be extracted from raw image observations via a *learned reward models*, which interestingly do introduce a series of high-impact, underex-

plored challenges in safety research in robot learning, similarly to their impact in LLMs Christian et al. (2025). Such autonomous robot systems might associate a "clean environment" (i.e., high-reward) with an absence of disturbances—a valid assumption in simulated environments where it is challenging to inject unpredictable human disturbances rather than one-step noise—and thus optimize for behaviors maintaining such immaculate state. When deployed in the real world, human movement in the scene might appear as just another form of unaccounted disturbance during training. If GM was to occur, the robot might enact actions resulting in blocking or restricting human presence in the room, as a means of having learned the spurious correlation between the absence of disturbances and clean state of the room. Such an issue only risks being intensified by the reliance on large-scale, end-to-end learning pipelines which prioritize performance over interpretability, risking catastrophic misalignment.

GM is an Inner Alignment Problem In the general framework of Inner-vs-Outer alignment (Hubinger et al., 2021) (Figure 1), GM primarily belongs to the class of inner alignment failures. Indeed, GM results from an inner alignment failure, due to the outer objective may be well-specified while being misrepresented internally by the agent. In practice, a robot could be perfectly rewarded for cleaning up but still learn an internal goal such as "remove objects from view", which could only correlate with "cleanliness" of the training scenarios. This subtle shift in representation makes the problem difficult to detect until the model encounters novel situations, and demonstrates that ensuring correct training signals is insufficient whenever internal generalization fails.

Why Does This Matter in Frontier AI? Robotics is currently advancing at an unprecedented rate, moving from industrial applications toward domestic and assistive environments. Modern systems are increasingly developed through the same *pre-training and fine-tuning paradigm* detailed in Ngo et al. (2025), with large robotics models adapted from general intelligence foundations—e.g., Abeyruwan et al. (2025) is adapted from Anil et al. (2025). Such convergence blurs the boundary between cognitive autonomy and physical embodiment, and amplifies the risks of misgeneralized objectives through the compounding effects of these two previously distinct domains.

2 Evaluation Design, Monitoring and Robustness

Assessing GM in autonomous systems requires moving beyond surface-level performance metrics—e.g., cumulative reward or *success rates*—toward interpreting the model’s internal representations and causal pathways connecting perception to action. In the PTT, this could mean identifying whether the rationale behind the robot’s actions originates from *detecting clutter* or by *seeking to preserve a static environment*. In practice, this connection can be constructed using techniques such as GradCAM (Selvaraju et al., 2020) for legacy convolutional models, or *circuits* for more modern transformer networks (Elhage et al., 2021). In particular, by analyzing activations corresponding to specific environmental cues, Elhage et al. (2021) show how to infer whether models do represent messiness semantically, merely associating it with the visual cues that correlate—albeit not immediately causally-related—to disorder, i.e., the presence of a human in a room.

Being a field in its relatively nascent stage, Robot Learning architectures have not been as thoroughly studied as their LLM counterpart, resulting in challenges related to the identification of said circuits. Nonetheless, advancements in research on language models proved the effectiveness of modeling transformers through circuits (Wang et al., 2022), and ultimately justify interest in developing similar techniques in the context of Robot Learning.

Safety via Interpretability In the context of interpretability *circuits* denote structured pathways of activations jointly encoding a concept (Elhage et al., 2021), which in the context of Robot Learning might equate to a specific fundamental behavior. Concretely, circuit discovery in Robot Learning can begin by analyzing video frames paired with internal representations to trace the emergence of task-specific signals. Since circuits are rarely studied in embodied systems, a circuit based approach would introduce a way to *visualize how goals are operationalized* in Robot Learning policies. For instance, discovering circuits could reveal whether the robot genuinely understands the notion of order or merely associates cleanliness with static visuals. This insight would enable more targeted interventions in both training and deployment phases, finding deception features and removing them to ensure the resulting system behaves in accordance with expectations. However, such evaluation system could still prove unsufficient to prevent GM due to the intricacies related to finding misaligned directions in the model’s activation space and.

Monitoring & Robustness Assessing the current status of autonomous system and its evolution revolves around developing interpretability-based evaluations that can truly capture the internal processes underpinning the performance of autonomous systems. In the context of GM in Robot Learning, continuous monitoring can involve periodically probing the model’s learned representations throughout training, verifying that its internal objectives remain aligned with external reward structures. Further, techniques such as activations *patching* (Elhage et al., 2021) and *probing* (Belinkov, 2021) can also reveal whether the circuits responsible for task execution retain stable, causal meaning or, more problematically, drift toward low performance induced by picking up on spurious correlations. *Red teaming*—either human (Feffer et al., 2024) or AI-based (Majumdar et al., 2025)—does also offer an important, complementary strategy to ensure robustness. By deliberately designing adversarial environments or synthetic disturbances that challenge the model’s learned assumptions, groups of model developers could stress-test whether the robot maintains intended behavior under targeted distributional shifts, building conviction towards the robustness of the algorithm considered.

Combining monitoring with adversarial testing provides a dual safeguard, as *interpretability assesses the inner workings of a model*, while *red-teaming verifies whether those goals remain safe and stable* across deployment scenarios. In practice, robustness could also be improved by incorporating *online monitoring probes* that run concurrently with the robot control system at eval and test-time, resulting in a high-level *semantic stop-button*, complementing red-teaming efforts while evaluating work deviating from the difference in intended behaviors that can be detected in order to correct or abort unsafe actions.

Limitations While circuits offer valuable insights into the model internals, complex behaviors which may violate the *Linear Representation Hypothesis* (Park et al., 2024) may defy traditional

interpretability assumptions. Additionally, interpretability constraints might even highlight forms of gaming and misuse, by exposing possible adversarial routes mimicking interpretable structures without genuine understanding. Thus, while circuits enhance transparency, their practical performance might improve when complemented with ongoing empirical validation and red teaming. Ideally, the practical implementation of autonomous systems in real-world scenarios should be based on a holistic safety strategy integrating interpretability, challenging behavioral evaluation, as well as rigorous governance oversight to ensure safety standards are met before and after advanced robot technologies enter society.

3 Governance and Policy

Just like the aviation and automotive industries, it is reasonable to expect the robotics industry to face regulatory scrutiny aiming at ensuring robots met security standards both *before* and *while* being integrated into society. Indeed, unlike purely digital systems, embodied AI operates in physical spaces that are shared with humans, so that failures can cause limited direct physical harm. Therefore, given their tangible risks and similarity with automotive and aviation precedents, it is rational to expect governmental institutions to leverage national jurisdiction to develop incentives for developers to adhere to strict safety standards, which could be enforced in practice by establishing comprehensive evaluation and interpretability requirements as a prerequisite for certification.

ICAO-like Governance for Decentralized and Unbalanced AI Regulation The *highly interconnected and international nature* of developments in the field of AI poses severe challenges to any single country in its ability to regulate such technology. Further, industry players' outsized influence in developing Frontier AI as part of a broader set of technologies permeating everyday-life does also pose severe *risks to the sovereignty of individual nation states* in addressing this matter, all of which resulting in clear governance challenges in addressing general AI development in an *internationally, non-coordinated* manner. A possible solution to this challenge would be the development of governance structures for (physical-) AI systems inspired from *international* regulatory bodies such as the International Civil Aviation Organization (ICAO), which harmonizes safety standards, certification procedures, and operational oversight *across nations*.

A comparable framework for rolling out robot learning in practice could consist on *shared safety benchmarks*, as well as *standardized interpretability requirements* and transparent reporting protocols among autonomous systems developers. Such an international alignment body could facilitate responsible scaling by ensuring that safety evaluations—including interpretability audits or robustness benchmarks—are conducted and validated under unified global standards, rather than fragmented national guidelines. This approach could foster cross-border accountability and mitigate the risk of a regulatory "race to the bottom", where looser jurisdictions attract unsafe or opaque deployments, by enforcing regulatory standards through the *positive network externalities* of agreeing on shared standards. Critically, however, in contrast to sectors like aviation, definitive inspections of models internals are rarely feasible at the present time, due the (currently) *fundamentally uninterpretable* nature of autonomous systems, whose internal logic cannot be easily dissected, exhaustively verified and thus audited. As a result of this fundamental technical limita-

tion, it is reasonable to expect oversight agencies might face fundamental uncertainties about the completeness and reliability of interpretability assessments, making post-certification monitoring and dynamic evaluation a critical complement to static regulation.

Considering the practical example of house robots, regulators could mandate interpretability as a crucial safety guarantee, requiring that deployed robots operate leveraging *demonstrable circuits* corresponding to key behavioral goals. By enforcing visibility into how decisions are made, authorities can reduce the likelihood of undetected GM. On a more technical note, mandating the adoption of behavioral cloning from human demonstrations instead of purely RL-based training may also offer a safeguard, as in such case human preferences would be directly *encoded through the demonstration data itself* rather than indirectly through reward optimization.

References

Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michael, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricu, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini Robotics: Bringing AI into the Physical World, March 2025.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricu, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis

Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gürä, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kociský, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturk, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivanee Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor To-

bin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemmingsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinker, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriiser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matač, Yadi Qian, Vikas Peswani, Paweł Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed,

Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Mau-rya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G. Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejas Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Mal-colm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yoge, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z. J. Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Nic-colò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappagantu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopal Vij, Konstantin Sha-gin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishabh Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luwei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael

Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeon Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan le, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, Mohammad-Hossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Cave-ness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaría-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolic-

chio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M, Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Fergn, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanou, Bo Feng, Keshav Dhandhana, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A Family of Highly Capable Multimodal Models, May 2025.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances, September 2021.

Francesco Capuano, Caroline Pascal, Adil Zouitine, Thomas Wolf, and Michel Aractingi. Robot Learning: A Tutorial, October 2025.

Brian Christian, Hannah Rose Kirk, Jessica A.F. Thompson, Christopher Summerfield, and Tsvetomira Dumbalska. Reward Model Interpretability via Optimal and Pessimal Tokens. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1048–1059, Athens Greece, June 2025. ACM. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732068.

Jonathan H. Connell and Sridhar Mahadevan, editors. *Robot Learning*. Springer US, Boston, MA, 1993. ISBN 978-1-4613-6396-5 978-1-4615-3184-5. doi: 10.1007/978-1-4615-3184-5.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Linxi "Jim" Fan. The Physical Turing Test: Jim Fan on Nvidia's Roadmap for Embodied AI, May 2025.

Michael Feffer, Anusha Sinha, Wesley Hanwen Deng, Zachary C. Lipton, and Hoda Heidari. Red-Teaming for Generative AI: Silver Bullet or Security Theater?, August 2024.

Ben Goertzel. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, December 2014. ISSN 1946-0163. doi: 10.2478/jagi-2014-0001.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems, December 2021.

Subhabrata Majumdar, Brian Pendleton, and Abhishek Gupta. Red Teaming AI Red Teaming, July 2025.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective, May 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024.

Ethan Perez, Sam Ringer, Kamilé Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny

Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, March 2017.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small, November 2022.