Expert Survey: Technical AI Safety & Security Research Priorities

Joe O'Brien [1 *] **Jeremy Dolan** [2 *] **Jay Kim** [3] **Jonah Dykhuizen** [2] **Jeba Sania** [4] **Sebastian Becker** [5]
Cara Labrador [1] **Jam Kraprayoon** [1]

Abstract

As AI systems approach broadly human-level performance, safety and security research is urgently needed to prevent severe harms from AI and ensure its benefits can be safely and reliably realized. To inform strategic investment, we surveyed 53 experts on the importance and tractability of 105 technical AI safety and security research areas. Each expert was asked to rate a subset of areas within their expertise. Our survey revealed several highly promising research directions centered around robust early warning and monitoring of AI risks. Some of the most promising sub-areas included specific capability evaluations (e.g., CBRN, cyber, and deception), understanding emergence and scaling laws, and advancing agent oversight. This study is the first to quantify expert priorities across a comprehensive taxonomy of AI safety and security research directions and to produce a data-driven ranking of their potential impact. These rankings may support evidence-based decisions about how to effectively deploy resources toward AI safety and security research.

1. Introduction

The development of increasingly capable general-purpose AI systems necessitates technical safety research to ensure that such systems will remain secure, reliable, and beneficial (Bengio et al., 2024; 2025b). This requires confronting a wide spectrum of technical challenges, from novel AI-specific concerns (such as alignment, control, and interpretability) to high-stakes versions of traditional cybersecurity and infrastructure threats (such as model theft, adversarial attacks, and infrastructure vulnerabilities).

[*]Equal contribution [1]Institute for AI Policy and Strategy (IAPS) [2]Independent Researcher [3]Williams College [4]Harvard Kennedy School [5]Effektiv Spenden. Correspondence to: Joe O'Brien <joe@iaps.ai>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

Despite widespread awareness of this urgent challenge (Leno da Silva et al., 2024; Bengio et al., 2025a), significant gaps remain in coordinating and prioritizing technical AI safety and security research efforts. Three bottlenecks have limited progress:

- **Resources:** Funding remains inadequate relative to the scale and urgency of the problem, and especially relative to the investment in capabilities.
- **Expertise:** AI safety and security faces a shortage of researchers with the necessary technical skills.
- **Uncertainty:** A lack of clarity about which research directions are most promising hinders effective prioritization by funders, policymakers, and researchers themselves.

Here we directly address the uncertainty bottleneck by surveying 53 experts from academia, industry, and civil society. Collectively, 105 sub-areas of technical AI safety and security research were assessed for *importance* (expected reduction in risk of severe harms) and *tractability* (ability for a marginal investment to make significant progress).

Our study reveals a consistent message from respondents: significant, actionable opportunities exist within technical AI safety and security research. 52 out of 53 respondents identified at least one research direction as both important and tractable (scoring ≥ 4.0 on both dimensions). More specifically, we reveal a critical immediate priority that emerged from our survey results: enhancing our ability to anticipate, detect, and monitor potentially harmful AI capabilities before they can cause widespread harm, through improving capability evaluations and evaluation science, understanding emergence and scaling laws, and advancing safety around agents and multi-agent systems.

2. Methodology

Our survey aimed to systematically capture experts' judgments of the importance and tractability of technical interventions related to AI safety and security.

2.1. Taxonomy Development

We developed a taxonomy of 105 technical AI safety and security research areas grouped into 20 high-level categories.

This taxonomy was constructed by distilling recent literature, e.g. Anwar et al. (2024) and Williams et al. (2024), into a mutually-exclusive taxonomy, and iterating that taxonomy several times through consultations with multiple experts to improve comprehensiveness. The focus was primarily on model-centric interventions—those with direct effects on AI model behavior, its evaluation, or its immediate operational environment. The full taxonomy is available in Appendix B.

2.2. Expert Recruitment & Demographics

We surveyed 53 experts, drawing from academia, industry, and civil society/non-profits, aiming for broad coverage across the areas of our taxonomy. 81% completed the optional demographic survey. A plurality of respondents were affiliated with academic institutions (39.5%), followed by non-profits (34.9%), with industry professionals comprising 16.3%, and smaller shares from government (2.3%), independent contributors (2.3%), and other affiliations (4.7%).

Most participants reported primary expertise in technical fields, with 65% in machine learning/AI and 19% in computer science/engineering. A small share specialized in public policy (7%) and social sciences (2%). The distribution of experience among participants was: 16% with over 10 years, 44% with 5 to 10 years, 30% with 3 to 5 years, and the remainder (10%) with less than 3 years.

Our recruitment strategy primarily involved systematically inviting first authors of exemplar publications in each subarea of our taxonomy, along with secondary authors possessing relevant expertise and experience. In total, 515 researchers were invited to participate.

The survey was administered anonymously through the Qualtrics platform from December 21, 2024 to March 4, 2025. Most experts were given a 10-week window to respond. During preliminary analysis, several categories were identified as having low response rates and an additional 45 experts working in those areas were identified and invited with a 9-day response window.

2.3. Survey Design & Metrics

Each expert was prompted to select top-level research categories corresponding with their expertise. Participants then rated sub-areas within those categories on two key dimensions using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), with an additional option to say "I don't know" for any sub-areas they were insufficiently familiar with. The assessed dimensions were:

- **Importance:** "Successfully addressing this issue would significantly reduce the risk of severe harm (loss of either >100 lives or >$10 billion USD in economic impact) from AI."
- **Tractability:** "An additional targeted investment of approximately $10M USD over the next two years would lead to significant, measurable advancements in addressing this issue."[1]

These definitions were designed to assess the potential for reduction in severe harms from an actionable level of funding ($10M USD) in order to assist decision makers considering strategic marginal investments.

From the importance and tractability ratings we then calculated a "promise score" (importance × tractability, max=25) which was used to rank each research area. Additionally, qualitative feedback was solicited for each high-level area on high-value challenges, missing sub-areas, and key obstacles.

3. Results & Discussion

Our results revealed clear patterns in research priorities. First and foremost, we identified directions where experts perceive significant potential for near-term impact with targeted investment ($10M USD over two years). We additionally analyze high-importance but low-tractability areas separately under "Strategic Long-Term Opportunities."

3.1. Most Promising Sub-Areas for Near-Term Progress

Areas with high promise scores represent directions where experts perceive significant potential for impactful advancements with targeted investment. The ten highest ranking sub-areas are listed in Table 1. Full results are available in Appendix A.

Table 1. Most Promising Sub-Areas

Sub-area	I	T	P	n
Emergence & task-specific scaling	5.00	4.25	**21.25**	4
CBRN evaluations	4.67	4.33	**20.22**	3
Evaluating deception, scheming	4.75	4.25	**20.19**	4
Oversight of LLM-agents	4.67	4.22	**19.70**	9
Cyber evaluations	4.50	4.25	**19.13**	4
Detecting unmeasured capabilities	4.38	4.25	**18.59**	16
Multi-agent metrics	4.43	4.14	**18.35**	7
Multi-agent security	4.57	4.00	**18.29**	7
Quantifying cyber threats	4.25	4.25	**18.06**	4
Managing value conflicts	4.50	4.00	**18.00**	4

I = Importance, T = Tractability, P = Promise (I×T),
n = Respondents for sub-area.

The dominant pattern among highly-ranked areas is a focus on evaluation, monitoring, forecasting, and detection of potentially dangerous capabilities. This includes understanding the relationship between emergent capabilities and scaling, evaluating specific misuse potentials (chemical, biological, radiological, and nuclear [CBRN], cyber, deception), and oversight of LLM-agents. The expert judgments here suggest that enhancing our capacity to understand and

[1]Note that this is $10M USD on the *current margin*, which implicitly reflects respondents' assessments of neglectedness.

anticipate what advanced AI systems can do and what risks they pose is more likely to succeed in the near-term than research on direct interventions to AI model safety and security.

Additionally, *multi-agent interactions* also emerged as a critical concern, with all sub-areas of this category ranking in the top 30% of priorities. This signals growing recognition that as AI systems become more agentic and interconnected, ensuring that their interactions yield safe outcomes is an increasingly urgent research frontier.

Experts consistently favored concrete, empirical approaches over theoretical frameworks, consistent with our near-term definitions for tractability. While more speculative research programs were acknowledged as important, they generally received lower tractability ratings compared to applied methodologies, such as evaluation. (For example, "Mechanistic understanding and limits of LLM reasoning" received mean ratings of 5.0 for importance but only 3.5 for tractability.) This suggests that for near-term impact, more practical work aimed at existing systems offers more leverage than developing general theories.

3.2. Strategic Long-Term Opportunities

Several research areas received a high rating for importance ($\geq$4.0) despite lower perceived tractability ($\leq$3.5). These domains represent strategic priorities that may require longer timelines and more substantial resource commitments.

Table 2. Importance–Tractability Gap

Sub-area	I	T	Gap	n
Access control & interface hardening	4.75	2.75	**2.00**	4
Supply chain integrity	4.57	3.00	**1.57**	7
Mechanistic understanding of LLMs	5.00	3.50	**1.50**	4
Preventing model self-exfiltration	4.00	2.75	**1.25**	4
Weight security & key management	4.50	3.25	**1.25**	4

I = Importance, T = Tractability, Gap = I−T, n = Respondents for sub-area.

Analysis of these areas reveals three key patterns: (1) Security implementation challenges dominate, with six of the ten largest importance-to-tractability gaps coming from cybersecurity domains; (2) These gaps appear primarily in applied areas rather than theoretical frameworks; and (3) Most require complex technical solutions spanning multiple domains. These findings point toward potential opportunities for larger, multi-year programs and public-private partnerships beyond the scope of typical grants.

3.3. Towards a Dual Investment Strategy

Overall, our findings point toward a dual investment strategy: prioritizing near-term resources for capability evaluations and evaluation science, understanding emergence and scaling laws, and advancing safety around agents and multi-agent systems, while simultaneously examining sustained, larger-scale investment in foundational areas like applied AI security (e.g., access control, supply chain integrity) and deep model understanding. These latter "moonshot" areas, though assessed as less tractable with a short time horizon, may yield dividends for long-term safety, with several rating $\geq$4.5 on importance despite lower tractability ratings.

Notably, our study revealed broad expert optimism about actionable opportunities in technical AI safety and security research. An overwhelming majority (52 of 53 experts) identified at least one research direction as both highly important and tractable (scoring $\geq$4.0 on both dimensions), signaling significant potential for progress with proper investment.

Our findings broadly corroborate other expert elicitations on promising technical research directions, e.g., Schuett et al. (2023), Grace et al. (2024), and Bengio et al. (2025a), lending further weight to the identified priority areas.

4. Policy Implications

Policy can be a powerful instrument in mitigating critical research gaps. Specifically, government can directly fund neglected areas, incentivize investment, play a coordination role, strengthen AI talent pipelines, and expand researcher access to frontier models.

- **Direct Funding:** Government's role in promoting research will depend on the source of the research gap. In some instances, research areas may be underfunded because of limited financial resources. To address this, Congress and relevant executive agencies, such as the National Science Foundation, should consider directly appropriating funding toward the most promising research areas identified in this report. Respondents classified areas as tractable if a $10M USD investment over the next two years would yield substantial advancements in AI safety and security—an unremarkable sum by government grantmaking standards that could yield significant safety dividends. In addition, government is uniquely positioned to direct higher sums to underfunded research areas that rate high on importance but low on tractability (e.g. supply chain security, access control and interface hardening), including through R&D agencies like the Defense Advanced Research Projects Agency (DARPA) or innovation-focused programs like the CHIPS and Science Act or the American Science Acceleration Project. It is unlikely that the AI R&D ecosystem will be able to address these gaps without the large, long-term investments usually provided by government.

- **Incentivizing Investment:** Beyond direct funding, government can: (1) List identified research gaps as

official priorities to signal interest; (2) Offer tax incentives or subsidies for AI safety and security research (Haykel, 2025); (3) Incorporate specific AI safety and security research commitments into broader agreements with industry stakeholders, and (4) Lower research costs by providing computational and other resources to safety researchers via structures like the National AI Research Resource (NAIRR) pilot.

- **Coordination:** Some research areas face neglect due to limited awareness rather than resource constraints. Government can address this coordination problem by proactively communicating research needs and establishing information-sharing mechanisms that provide visibility into AI R&D progress.

- **Talent Development:** Underinvestment often stems from talent shortages. Short-term solutions include scholarships for critical AI safety and security areas and pathways for international talent, while long-term efforts should expand the domestic AI talent pipeline through educational funding.

- **Researcher Access:** Limited access to advanced AI systems hampers safety research. Policymakers can democratize access by encouraging industry to waive costs for under-resourced organizations and create mechanisms like regulatory sandboxes for government researchers to access models pre-deployment.

5. Limitations & Future Work

Our study has important limitations: modest sample size (N=53), uneven coverage across the taxonomy (31 sub-areas were cut from analysis due to not meeting our minimum of 3 responses), and its nature as a static snapshot in a rapidly evolving field. The high percentage of experts with technical backgrounds may have skewed results away from topics such as fairness or privacy. The survey's anonymous nature made it impossible to verify whether experts favored their own research areas, although we invited experts from all subareas to mitigate and ideally average out this bias.

Results should be interpreted as directional indicators rather than definitive priorities. Future work should pursue broader sampling with higher response rates, explore expert rationales through qualitative follow-up studies, and develop real-time elicitation methods to capture shifting priorities in this dynamic domain.

6. Conclusion

Navigating the complex landscape of AI safety and security research demands clear prioritization as human-level capabilities rapidly develop. Our survey of 53 experts across 105 technical research directions provides data-driven guidance for funders, researchers, and policymakers.

Expert assessments revealed the immediate priority lies in enhancing early warning and monitoring by improving capability evaluations and evaluation science, understanding emergence and scaling laws, and advancing safety around agents and multi-agent systems. The consistent prioritization of research enabling strategic visibility over research aimed at direct risk-reduction underscores a critical insight: managing the risks of human-level AI will require robust systems to understand and anticipate what advanced AI systems can do.

Our findings also point towards strategic long-term investments—areas like security engineering and deep model understanding that require larger-scale, coordinated efforts despite lower near-term tractability. This suggests a dual investment strategy: addressing urgent visibility needs while building foundational capabilities for complex, long-horizon problems.

Experts consistently favored concrete, empirical interventions over theoretical frameworks. While AI companies are active in some top-ranked areas, opportunities exist for targeted support of independent research that provides impartial assessment of safety measures.

Despite limitations in sample size, this work provides a quantitative snapshot of expert opinion and serves as a template for iterative improvement in guiding AI safety and security research prioritization.

Acknowledgements

Assessing the landscape of AI safety and security is an inherently interdisciplinary and ambitious endeavor, made possible only through generous collaboration. We extend our sincere gratitude to the following individuals whose insights and thoughtful feedback greatly enhanced the quality and rigor of this work: Usman Anwar, Richard Ren, Peter Barnett, Zoe Williams, Ari Holtzman, Oscar Delaney, Micah Carroll, Buck Shlegeris, Daniel Kang, Peter Hase, Daniel Brown, Nouha Dziri, Chirag Agarwal, Niloofar Mireshghallah, Leilani Gilpin, David Krueger, Noemi Dreksler, Willem Sleegers, and David Moss. We are also grateful to our anonymous reviewers for their useful feedback and suggestions, and, of course, to all of our anonymous survey respondents. This research was conducted in part through the Supervised Program for Alignment Research; we thank the program organizers for providing the support and intellectual environment that made this research possible. All remaining errors are our own.

Impact Statement

This paper presents an expert-driven prioritization of AI safety and security research, designed to guide resource allocation and enhance society's capacity to anticipate, understand, and prepare for rapidly advancing AI systems. Our primary goal is to accelerate the development of robust safety measures, thereby fostering the realization of AI's considerable societal potential. We offer these findings as a data-driven tool to empower funders, policymakers, and the research community in making more informed strategic decisions within the complex safety and security landscape. Applying these findings responsibly is crucial; we encourage readers to engage with them critically, as one valuable input alongside diverse research perspectives, taking into account this study's specific context, the field's dynamism, and the nuanced execution that some identified high-impact research areas (like dangerous capability evaluations) inherently demand. Ultimately, this work seeks to contribute to a more effective, coordinated global ecosystem striving for the development of advanced AI that is demonstrably safe, secure, and aligned with human interests and values.

References

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., Korbak, T., Zhang, H., Zhong, R., hÉigeartaigh, S. O., Recchia, G., Corsi, G., Chan, A., Anderljung, M., Edwards, L., Petrov, A., Witt, C. S. d., Motwan, S. R., Bengio, Y., Chen, D., Torr, P. H. S., Albanie, S., Maharaj, T., Foerster, J., Tramer, F., He, H., Kasirzadeh, A., Choi, Y., and Krueger, D. Foundational challenges in assuring alignment and safety of large language models, September 2024. URL https://arxiv.org/abs/2404.09932.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. Managing extreme ai risks amid rapid progress. *Science*, 384 (6698):842–845, May 2024. ISSN 1095-9203. doi: 10.1126/science.adn0117. URL http://dx.doi.org/10.1126/science.adn0117.

Bengio, Y., Maharaj, T., Ong, L., Russell, S., Song, D., Tegmark, M., Xue, L., Zhang, Y.-Q., Casper, S., Lee, W. S., Mindermann, S., Wilfred, V., Balachandran, V., Barez, F., Belinsky, M., Bello, I., Bourgon, M., Brakel, M., Campos, S., Cass-Beggs, D., Chen, J., Chowdhury, R., Seah, K. C., Clune, J., Dai, J., Delaborde, A., Dziri, N., Eiras, F., Engels, J., Fan, J., Gleave, A., Goodman, N., Heide, F., Heidecke, J., Hendrycks, D., Hodes, C., Hsiang, B. L. K., Huang, M., Jawhar, S., Jingyu, W., Kalai, A. T., Kamphuis, M., Kankanhalli, M., Kantamneni, S., Kirk, M. B., Kwa, T., Ladish, J., Lam, K.-Y., Sie, W. L., Lee, T., Li, X., Liu, J., Lu, C., Mai, Y., Mallah, R., Michael, J., Moës, N., Möller, S., Nam, K., Ng, K. Y., Nitzberg, M., Nushi, B., hÉigeartaigh, S. O., Ortega, A., Peigné, P., Petrie, J., Prud'Homme, B., Rabbany, R., Sanchez-Pi, N., Schwettmann, S., Shlegeris, B., Siddiqui, S., Sinha, A., Soto, M., Tan, C., Ting, D., Tjhi, W., Trager, R., Tse, B., H., A. T. K., Wilfred, V., Willes, J., Wong, D., Xu, W., Xu, R., Zeng, Y., Zhang, H., and Žikelić, D. The Singapore Consensus on Global AI Safety Research Priorities, May 2025a. URL https://aisafetypriorities.org/files/Singapore_Consensus_2025.pdf.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Ajala, O., Albalawi, F., Alserkal, M., Avrin, G., Busch, C., de Carvalho, A. C. P. d. L. F., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Johnson, C., Jolly, G., Katzir, Z., Khan, S. M., Kitano, H., Krüger, A., Lee, K. M., Ligot, D. V., López Portillo, J. R., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Pezoa Rivera, R., Ravindran, B., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. International AI Safety Report. Technical Report DSIT 2025/001, Department for Science, Innovation and Technology and AI Safety Institute, January 2025b. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., and Brauner, J. Thousands of AI authors on the future of AI, April 2024. URL https://arxiv.org/abs/2401.02843.

Haykel, I. AI security tax incentives. Policy brief from Americans for Responsible Innovation, March

2025. URL https://ari.us/policy-bytes/ai-security-tax-incentives.

Leno da Silva, F., Glatt, R., Giera, B., Gonzales, C., Bremer, P.-T., Newman, J., Corley, C., Stracuzzi, D., Kegelmeyer, P., Alexander, F. J., Gal, Y., Greaves, M., Gleave, A., Lillicrap, T., Falet, J.-P., and Bengio, Y. Safety in artificial intelligence: Challenges and opportunities for the U.S. national labs and beyond. Technical Report LLNL-TR-2000782, Lawrence Livermore National Laboratory, 2024. URL https://data-science.llnl.gov/safety-artificial-intelligence-challenges-opportunities-us-national-labs-beyond.

Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., and Garfinkel, B. Towards best practices in AGI safety and governance: A survey of expert opinion, May 2023. URL https://arxiv.org/abs/2305.07153.

Williams, J., Rönn, K., Graabak, J., Sirin, B., Weiler, S., Gonzales-Puell, E., Reistad, E., Schoneveld, K., Duhamel, C., Tan, C., Celik, B., and Depuydt, M.-A. Risk and reward: AI assurance technology market report, May 2024. URL https://www.aiat.report/.

A. Sub-Areas Ranked by Promise Score

The following table contains all ratings from our survey ordered by promise score (importance × tractability). Sub-areas that received two or fewer ratings in either importance or tractability were excluded from quantitative analysis due to insufficient data for meaningful statistical interpretation, which led to the exclusion of 31 sub-areas.

Sub-area	I	T	P	n
Emergence and task-specific scaling patterns	5.00	4.25	21.25	4
CBRN (Chemical, Biological, Radiological, and Nuclear) evaluations	4.67	4.33	20.22	3
Evaluating deception, scheming, situational awareness, and persuasion	4.75	4.25	20.19	4
Oversight and monitoring of LLM-agents	4.67	4.22	19.70	9
Cyber evaluations	4.50	4.25	19.13	4
Detecting and addressing previously unmeasured or latent capabilities	4.38	4.25	18.59	16
Multi-agent metrics and evaluations	4.43	4.14	18.35	7
Multi-agent security	4.57	4.00	18.29	7
Quantifying cyber threats from advanced capabilities	4.25	4.25	18.06	4
Manage conflicts between different values	4.50	4.00	18.00	4
Safety and emergent functionality in multi-agent interactions	4.57	3.86	17.63	7
Validating and applying interpretability methods	4.38	4.00	17.50	8
Mechanistic understanding and limits of LLM reasoning	5.00	3.50	17.50	4
Evaluation methodology and metrics	4.14	4.20	17.40	15
Control mechanisms for untrusted models	4.00	4.25	17.00	4
Transparency, information asymmetries, and communication protocols	4.00	4.25	17.00	5
Detecting and preventing collusion and emergent collective behaviour	4.43	3.80	16.83	7
Network effects and destabilizing dynamics in agent ecosystems	4.33	3.83	16.61	6
Pluralistic value alignment	4.00	4.00	16.00	4
Evaluating tool affordances for LLM-agents	4.11	3.89	15.99	9
Develop more robust evaluations for which values an LLM encodes	3.75	4.25	15.94	4
Studying misalignment through simplified model organisms	4.07	3.86	15.70	14
Confidential computing and environment isolation	4.57	3.43	15.67	7
Theoretical foundations for evaluation	4.07	3.81	15.50	16
Improving evaluation robustness	3.80	4.07	15.45	15
Robustness to underspecification	4.17	3.67	15.28	6
Pretraining alterations to improve Interpretability	4.00	3.80	15.20	5
Understanding how finetuning changes a pretrained model	4.00	3.75	15.00	5
Transparency	4.13	3.63	14.95	8
Foundational research on operationalizing values in LLMs	4.25	3.50	14.88	4
Weight security and key management	4.50	3.25	14.63	4
Debate	4.10	3.50	14.35	10
Continuous monitoring, advanced threat detection, and incident response	3.86	3.71	14.33	7
Datacenter security	4.14	3.43	14.20	7
Uncertainty quantification	3.75	3.75	14.06	4
Defending against poisoning and backdoors	3.75	3.75	14.06	4
Justify value choices for alignment	4.00	3.50	14.00	4
Scalable data auditing, filtering, and Pretraining with Human Feedback (PHF)	4.00	3.50	14.00	4
Eliciting Latent Knowledge (ELK)	4.25	3.25	13.81	8
Supply chain integrity and secure development	4.57	3.00	13.71	7
Explainability	4.00	3.43	13.71	7
Develop output-based adversarial training techniques for more robust alignment	3.80	3.60	13.68	5
Better elicitation mechanisms from humans	3.60	3.67	13.20	10
Access control and interface hardening	4.75	2.75	13.06	4
Building verifiable and robust AI architectures	3.92	3.33	13.06	12
Lifelong learning and goal-directedness in LLM agents	3.67	3.56	13.04	9
Iterated Distillation and Amplification (IDA)	4.00	3.22	12.89	9
Tamper-evidence and tamper-proofing	3.57	3.57	12.76	7
Hardware-integrated monitoring and verification	3.57	3.50	12.50	7
Developmental Interpretability	3.75	3.29	12.32	8
Defending against jailbreaks and prompt injections	3.75	3.25	12.19	4
Detecting modified models or poisoned data	4.00	3.00	12.00	4
Decision theory and rational agency	3.70	3.20	11.84	10
Formal verification of AI systems	3.82	3.09	11.80	11
Interpretability foundations	3.38	3.43	11.57	8
Peer incentivisation and automated mechanism design	3.00	3.75	11.25	4
Human-AI Interaction and collaboration	3.20	3.50	11.20	5
Preventing model self-exfiltration	4.00	2.75	11.00	4
Embedded agency	3.55	3.09	10.96	11
Feature and circuit analysis	3.43	3.17	10.86	7
Specialized chips to compute encrypted data	3.57	3.00	10.71	7
Limiting models' ability to perform harmful tasks	3.60	2.80	10.08	5
Recursive Reward Modeling	3.22	3.13	10.07	9
Mechanism design and multi-agent communication	2.80	3.50	9.80	5
Retrieval-augmented pre-training	2.60	3.75	9.75	5
Adversarial robustness to perturbations	3.25	3.00	9.75	4
Scalable techniques for targeted modifications of LLM behavior	3.00	3.00	9.00	5
Theoretical foundations of deep learning	2.67	3.33	8.89	3
Reinforcement Learning from AI Feedback (RLAIF)	3.10	2.80	8.68	10

Continued on next page

Table 3: Sub-areas Ranked by Promise Score (continued)

Sub-area	I	T	P	n
Limits of Transformers	2.33	3.67	8.56	3
Causal incentives	2.90	2.90	8.41	10
Control theory applications in AI safety	3.09	2.60	8.04	11
Model robustness and oracle protection	3.00	2.67	8.00	4
Safe reinforcement learning for non-LLM systems	3.00	2.50	7.50	3
Double descent and overparameterization	2.00	3.33	6.67	3
Implicit bias of optimization algorithms	2.67	2.33	6.22	3
Optimization and loss landscape analysis	2.00	2.33	4.67	3

I = Importance, T = Tractability, P = Promise (I×T, max=25),
n = Number of respondents. (Subareas with two or fewer respondents were excluded from results.)

B. Taxonomy

Top-level categories

Respondents first selected one or more of the following categories:

1. Theoretical foundations and provable safety in AI systems
2. Training and finetuning methods for alignment and safety
3. Scalable oversight and alignment techniques
4. Understanding in-context learning, reasoning, and scaling behavior
5. Interpretability, explainability, and transparency
6. Robustness
7. Improving the science of AI evaluation
8. Domain-specific AI evaluation design
9. Agentic LLMs and single-agent risks
10. Multi-agent interactions
11. Cooperative AI and mechanism design
12. Fairness
13. Accountability
14. Ethics
15. Choosing and operationalizing values in AI
16. Privacy
17. Cybersecurity for AI models
18. Hardware and infrastructure security for AI
19. Improving general understanding of deep learning
20. Research on safety in non-LLM systems

Category and sub-area descriptions

The following area descriptions were then shown to respondents alongside importance and tractability questions.

1. **Theoretical foundations and provable safety in AI systems:** Advancing the theoretical foundations of AI safety by building models and frameworks that ensure provably correct and robust behavior. These efforts span from verifiable architectures and formal verification methods to embedded agency, decision theory, incentive structures aligned with causal reasoning, and control theory.
    a. **Building verifiable and robust AI architectures:** Constructing AI systems with architectures that support formal verification and robustness guarantees, such as world models that enable safe and reliable planning, or guaranteed safe AI with Bayesian oracles. This area emphasizes simplicity and transparency to aid in provability.
    b. **Formal verification of AI systems:** Applying formal methods to verify that AI models and algorithms meet stringent safety, robustness, and performance criteria. This includes proving resilience against adversarial inputs and perturbations, and certifying conformance to specified safety properties under varying conditions.
    c. **Decision theory and rational agency:** Establishing formal decision-making frameworks that ensure rational and safe choices by AI agents, potentially drawing on concepts like causal and evidential decision theory.
    d. **Embedded agency:** Explores how agents can model and reason about themselves and their environment as interconnected parts of a single system, addressing challenges like self-reference, resource constraints, and the stability of reasoning processes. This includes tackling problems arising from the lack of a clear boundary between the agent and its environment.
    e. **Causal incentives:** Developing frameworks that formalize how to align agent incentives with safe and desired outcomes by ensuring their causal understanding matches intended objectives. This research provides a formal language for guaranteeing safety, addressing challenges like goal misspecification, and complementing broader efforts in agent foundations and robust system design.
    f. **Control theory applications in AI safety:** Leveraging principles from control theory to ensure stability, robustness, and safety for AI-driven systems interacting with dynamic physical environments. This includes designing controllers and feedback mechanisms to maintain system integrity, prevent runaway behaviors, and achieve desired performance criteria under uncertainty.

2. **Training and finetuning methods for alignment and safety:** Developing reliable training and finetuning strategies for AI models to ensure that their outputs remain safe, interpretable, and aligned with intended goals. This involves understanding how finetuning affects model behavior, employing adversarial training for robust alignment, carefully adjusting pre-training processes, and improving data quality and auditing methods.
    a. **Understanding how finetuning changes a pretrained model:** Investigating how finetuning alters a model's internal representations and behaviors to better predict, and ultimately control, downstream safety outcomes.
    b. **Develop output-based adversarial training techniques for more robust alignment:** Developing training procedures, such as adversarial training focused on internal model representations, or 'process supervision,' that directly optimize against adversarial examples and undesirable outputs, making models more resistant to manipulations that could lead to unsafe behaviors.
    c. **Scalable techniques for targeted modifications of LLM behavior (including unlearning):** Creating scalable methods for precisely adjusting model outputs, such as removing unwanted content or refining responses to adhere to alignment constraints without broadly degrading performance. This may also include removal of *unknown* or latent undesirable capabilities that emerge in large models.
    d. **Retrieval-augmented pre-training:** Incorporating retrieval mechanisms during pre-training to better ground models in verified information.
    e. **Pretraining alterations to improve interpretability:** Altering pre-training protocols to produce models with clearer internal representations and decision-making pathways, allowing for more effective downstream analysis and intervention.
    f. **Limiting models' ability to perform harmful tasks:** Introducing mechanisms during pre-training that proactively limit a model's potential to learn or perform harmful tasks, constraining the model's capability space to safer domains before downstream fine-tuning.
    g. **Scalable data auditing, filtering, and Pretraining with Human Feedback (PHF):** Developing tools for large-scale data auditing, filtering, training-data attribution, and incorporating human feedback at the pre-training stage.

3. **Scalable oversight and alignment techniques:** Developing approaches to guide and align increasingly complex AI systems even in tasks where direct oversight is challenging, such as by the use of AI feedback, debate, iterative training processes, and enhanced elicitation methods.
    a. **Reinforcement Learning from AI Feedback (RLAIF):** Using feedback generated by AI systems to guide reinforcement learning, effectively scaling the oversight process beyond purely human-labeled data.
    b. **Debate:** Encouraging multiple models (or model instances) to discuss and critique each other's reasoning, with human overseers judging the best arguments.
    c. **Iterated Distillation and Amplification (IDA):** An alignment approach where increasingly capable AI systems are trained by recursively using weaker AIs to teach and amplify smarter successors. To address the limitations of human-defined feedback and reward functions, IDA decomposes complex tasks—using AI assistance—into simpler subtasks with accessible human or algorithmic evaluation signals, enabling scalable alignment and improved performance over time.
    d. **Better elicitation mechanisms from humans:** Improving methods to extract more reflective, aspirational, and consistent human preferences, to provide data to guide AI systems along these preferences and update in accordance with changes in values over time.
    e. **Recursive reward modeling:** Breaking down complex tasks into simpler subtasks for which reward signals can be more easily specified, then "building up" to oversee more complex behaviors.

4. **Understanding in-context learning, reasoning, and scaling behavior:** Methods to gain a comprehensive understanding of how large language models learn, reason, and scale, such as by examining in-context learning (ICL) mechanisms, the influence of data and design on behavior, the theoretical foundations of scaling, the emergence of advanced capabilities, and the nature of reasoning.
    a. **Mechanistic understanding of In-Context Learning:** Investigating the internal processes by which transformers perform ICL, including whether these processes resemble emergent optimization behavior, advanced pattern-matching, or other structural mechanisms. This research may include scenario-based analyses to identify the circuits critical for ICL under artificial constraints.

b. **Influences on ICL behavior and performance:** Examining how the tasks, instructions, pre-training data distribution, and design choices (e.g., instruction tuning, model size, training duration) shape the range and reliability of behaviors that can be specified in-context.

c. **Theoretical and representational aspects of scaling:** Clarifying when and how scaling drives improvements, such as by building a more robust theoretical framework to describe scaling laws, or analyzing how increasing model size and training data influence learned representations.

d. **Emergence and task-specific scaling patterns:** Formalizing and forecasting the emergence of new capabilities as models scale, investigating whether scaling alone can produce certain capabilities, and designing methods for discovering task-specific scaling laws.

e. **Impact of scaling and training on reasoning capabilities:** Determining whether and how increases in model size and training complexity enhance reasoning abilities, and identifying which aspects of training conditions and data sources facilitate the acquisition of reasoning skills.

f. **Mechanistic understanding and limits of LLM reasoning:** Examining the underlying mechanisms of reasoning in LLMs, exploring non-deductive reasoning capabilities of LLMs (e.g., causal or social reasoning).

g. **Limits of Transformers:** Defining the computational limits of transformers in supporting sophisticated reasoning.

5. **Interpretability, explainability, and transparency:** Ensuring that AI systems are understandable, trustworthy, and transparent. This involves developing tools and methods to interpret model internals, refining the reliability and scalability of interpretability techniques, exploring ways to elicit and explain model reasoning, and improving the transparency of complex models.

   a. **Interpretability foundations:** Focuses on theoretical and experimental studies investigating how models represent and encode concepts, emphasizing structural and abstraction-level insights, including by distinguishing linear from non-linear encodings, understanding polysemanticity and superposition, examining concept mismatches between models and humans, and discovering more accurate abstractions for interpretability.

   b. **Validating and applying interpretability methods:** Developing rigorous criteria and benchmarks for evaluating the reliability of interpretability methods, and understanding whether these methods maintain their validity when applied to actively modify model behavior.

   c. **Feature and circuit analysis:** Creating scalable approaches for feature interpretation, circuit discovery, and feature steering (or top-down/control vectors).

   d. **Eliciting Latent Knowledge (ELK):** Developing methods to reveal hidden knowledge embedded within models, enabling researchers to identify what models implicitly "know" about the world and how this knowledge influences predictions.

   e. **Developmental interpretability:** Investigating how AI models' internal representations and behaviors evolve throughout their training process to understand the developmental stages and mechanisms by which complex capabilities emerge. This research aims to uncover the progressive changes in model structure and function, facilitating better alignment and safety assurances.

   f. **Transparency.** Research here aims to open up the black box of AI systems by uncovering how data, architecture, and training processes shape model outputs. Example research focuses on advanced documentation frameworks, auditing tools to surface biases or vulnerabilities, and reporting protocols to effectively explain outputs and communicate uncertainty.

   g. **Explainability:** Methods to understand why a model generates specific outputs. Technical approaches include developing post-hoc or embedded explanation methods, measuring and improving explanation fidelity, and crafting user-focused interfaces that clarify causal or logical relationships.

6. **Robustness:** Ensuring that AI systems remain reliable and secure in the face of adversarial manipulation, misaligned inputs, and uncertain conditions, such as by protecting against prompt-based exploits, poisoning attacks, and adversarial perturbations, and introducing control mechanisms and uncertainty quantification methods to maintain resilient system behavior at scale.

   a. **Defending against jailbreaks and prompt injections:** Improving state-of-the-art methods for discovering, evaluating, and defending against prompt injection and "jailbreaking" attacks. Research also focuses on structural defenses, such as detection, filtering, and paraphrasing of prompts, as well as addressing vulnerabilities stemming from a lack of robust privilege levels (e.g., system prompt vs. user instruction) in LLM inputs.

b. **Defending against poisoning and backdoors:** Understanding how LLMs can be compromised through data poisoning at various training stages, examining the effect of model scale on vulnerability, testing out-of-context reasoning under poisoning, and exploring attacks via additional modalities and encodings. This area also includes detecting and removing backdoors (i.e., Trojan detection) to ensure that covertly embedded harmful behaviors are mitigated.

c. **Adversarial robustness to perturbations:** This area investigates how models can be made more resilient to carefully crafted adversarial perturbations designed to degrade performance or reveal vulnerabilities. Research involves identifying methods for bolstering model robustness under challenging conditions, including adversarial training and certified defenses.

d. **Control mechanisms for untrusted models:** Designing and evaluating protocols to control outputs from untrusted models. This includes methods for monitoring backdoored outputs, integrating control measures with traditional insider risk management strategies, building safety cases for control tools, and employing white-box techniques (e.g., linear probes) for continuous oversight.

e. **Uncertainty quantification:** Quantifying uncertainty in model predictions. Techniques include ensemble methods, conformal predictions, and Bayesian approaches to estimate and calibrate model confidence.

7. **Improving the science of AI evaluation:** Ensuring that AI systems can be accurately assessed and understood. This includes theoretical work in capability and safety evaluation, as well as improving the reliability and fairness of evaluation processes.

   a. **Theoretical foundations for evaluation:** Research includes creating rigorous frameworks for predicting capabilities (as opposed to relying solely on benchmarks) understanding generality and generalization in LLMs, and developing theory-grounded taxonomies of model capabilities.

   b. **Evaluation methodology and metrics:** Research focuses on designing holistic, theory-grounded metrics (e.g. focused on more than just harmlessness), accounting for scaffolding in evaluations, and characterizing safety-performance trade-offs.

   c. **Studying misalignment through simplified model organisms:** Developing and studying simplified AI models—"model organisms"—to probe potential misalignment, gain insights into failure modes, and refine evaluation strategies without the complexity of full-scale systems.

   d. **Improving evaluation robustness:** Methods here aim to stabilize evaluations against sensitivity to prompts, detect and address contaminated data, ensure that evaluations remain meaningful even if models are fine-tuned in a targeted manner for certain tasks, and mitigate bias in AI evaluations (including biases in crowdsourced human evaluations).

   e. **Detecting and addressing previously unmeasured or latent capabilities:** Developing strategies to uncover latent harmful abilities within AI models and prevent models from exhibiting undesirable behaviors such as "sandbagging" or deceptively underperforming during evaluations.

8. **Domain-specific AI evaluation design:** Developing specialized evaluation tools to assess AI models' capabilities and safety in critical areas such as automated AI research and development, cybersecurity, chemical/biological/radiological/nuclear (CBRN) scenarios, and manipulative behaviors like deception and persuasion.

   a. **Automated AI R&D evaluations:** Designing evaluations to assess a model's capacity to generate research ideas, propose improvements to algorithms, or autonomously advance AI capabilities.

   b. **Cyber evaluations:** Designing evaluations to assess a model's ability to understand, exploit, or defend against cybersecurity threats and vulnerabilities.

   c. **CBRN (Chemical, Biological, Radiological, and Nuclear) evaluations:** Designing evaluations to assess a model's understanding of hazardous CBRN materials and scenarios, ensuring it cannot be easily leveraged to facilitate harmful acts involving these agents.

   d. **Evaluating deception, scheming, situational awareness, and persuasion:** Designing evaluations to assess how well models can deceive, strategize, maintain situational awareness, or influence human decision-making.

9. **Agentic LLMs and single-agent risks:** Developing a deeper understanding of agentic behavior in LLM-based systems. This work clarifies how LLM-agents learn over time, respond to underspecified goals, and engage with their environments.

a. **Lifelong learning and goal-directedness in LLM agents:** Investigating how agentic LLMs evolve through ongoing learning, and potentially exhibit undesirable behaviors due to goal-directedness.
b. **Robustness to underspecification:** Enhancing methods to ensure LLM-agents remain aligned despite vague or shifting objectives.
c. **Oversight and monitoring of LLM-agents:** Building automated oversight and monitoring tools to track LLM-agent actions.
d. **Evaluating tool affordances for LLM-agents:** Evaluating the safety of providing LLM-agents with tools and affordances, and determining whether robust safety assurances are possible for given affordances.

10. **Multi-agent interactions:** Research focusing on ensuring safe multi-agent interactions, such as by detecting and preventing malicious collective behaviors, studying how transparency can affect agent interactions, and developing evaluations for agent behavior and interaction.

a. **Safety and emergent functionality in multi-agent interactions:** Understanding how individual agent dispositions and capabilities scale into complex multi-agent dynamics, evaluating emergent functionalities (e.g., coordinated strategies), enhancing robustness of LLM agents to correlated failures stemming from foundationality, and applying insights from multi-agent RL research to LLM-based systems.
b. **Detecting and preventing collusion and emergent collective behavior:** Developing detection techniques (e.g., information-theoretic or interpretability-based) for collusion between AI agents, benchmarking and evaluating collusive tendencies, designing mitigation strategies such as oversight regimes, communication restrictions, and methods for steering agents, understanding conditions (e.g., agent similarity, communication channels, environment structure) that facilitate collusion, and understanding why and how general "super-agents" might develop from many narrow agents.
c. **Multi-agent security:** Assessing unique security risks that arise in multi-agent ecosystems, designing defenses (e.g., secure communication protocols, improved network architectures, information security), studying how multiple systems can circumvent safeguards, evaluating robustness of cooperation to adversarial attacks (e.g., if a small number of malicious agents can destabilize larger groups), evaluating how well agents can adversarially attack each other, and studying the impact of AI agent's training dynamics on data generated by each other with respect to shared vulnerabilities/correlated failure modes.
d. **Network effects and destabilizing dynamics in agent ecosystems:** Understanding which network structures and interaction patterns lead to robust or fragile systems, monitoring and controlling dynamics and co-adaptation of networks of advanced agents, and identifying important security concerns in existing and future multi-agent application areas (e.g., finance, energy grids) and applying lessons from those areas to manage destabilizing forces.
e. **Transparency, information asymmetries, and communication protocols:** Studying how agent transparency (e.g., code access) or predictability of agents can influence cooperation or defection, scaling Bayesian persuasion and information design to complex multi-agent settings, developing secure information transmission methods between AI agents to promote cooperation, examining how agent similarity and evidential reasoning about others affect ability and propensity to cooperate, and developing efficient algorithms for zero- or few-shot coordination in high-stakes scenarios.
f. **Multi-agent metrics and evaluations:** Distinguishing and measuring cooperative dispositions, understanding agents' robustness against coercion or exploitation, quantifying traits like altruism or spite, assessing the impact of capability asymmetries between agents, examining how training processes and data sources influence cooperation, and developing dangerous capability evaluations for multi-agent systems.

11. **Cooperative AI and mechanism design:** Fostering beneficial multi-agent ecosystems through research on human-AI interaction, mechanism design, communication protocols, peer incentivization, and automated mechanism design.

a. **Human-AI interaction and collaboration:** Designing AI systems that can understand and predict human actions and preferences; creating interfaces and protocols for effective human-AI teamwork; understanding how interactive AI may change human decision making.
b. **Mechanism design and multi-agent communication:** Focuses on foundational concepts like social choice theory, incentive alignment, and emergent communication protocols in multi-agent systems to ensure cooperation and fair outcomes.

c. **Peer incentivisation and automated mechanism design:** Focuses on practical and scalable applications of mechanism design, including methods for incentivizing cooperation among agents, designing secure and scalable inter-agent contracting and norm enforcement mechanisms, and structured opponent-shaping strategies in complex environments.

12. **Fairness:** Research focusing on developing equitable AI systems, including detecting and mitigating bias, ensuring fair representation across diverse groups, addressing fairness in dynamic or constrained data scenarios, and reconciling conflicting fairness definitions to align interventions with societal values.

    a. **Fairness under dynamic and constrained data scenarios:** Ensuring that fairness interventions remain effective under continual learning, adaptive deployment, or evolving operational contexts.
    b. **Fair representation and participation in AI systems:** Promoting fair representation and generalization across different subpopulations, and ensuring inclusive participation in the development and governance of AI systems.
    c. **Bias detection, quantification, and mitigation techniques:** Developing systematic methods to detect, measure, and reduce bias in model outputs, ranging from pre-processing adjustments to post-hoc corrections. This may also include causal methods for fairness, such as causal modeling techniques to distinguish between genuine causal relationships and spurious correlations in observed disparities, enabling fairness interventions that address underlying structural causes.
    d. **Fairness in multilingual, cross-cultural, and multimodal contexts:** Addressing fairness challenges that arise when models operate across different languages, cultures, and data modalities.
    e. **Intersectional fairness and complex group structures:** Addressing compounded biases that arise when protected attributes overlap, such as race and gender, to ensure fairness approaches capture nuanced harms across intersectional groups. This research develops computational methods and evaluation frameworks to avoid oversimplifying population categories and to identify disparities affecting complex group structures.
    f. **Reconciling multiple fairness definitions and normative trade-offs:** Comparing and combining conflicting formal definitions of fairness to address the normative trade-offs they entail and align fairness interventions with societal values. This research clarifies the theoretical and practical implications of fairness definitions, helping practitioners navigate complex policy and ethical considerations.

13. **Accountability:** Research focusing on ensuring AI systems are transparent, reliable, and compliant, including developing auditing tools, attributing AI outputs to specific models, mitigating risks of power concentration in AI development and deployment, and automating regulatory compliance.

    a. **Auditing mechanisms:** Developing automated post-hoc auditing tools, using privacy-enhancing technology to facilitate secure audit access to sensitive data, building auditability into systems by design, and establishing continuous accountability pipelines that monitor, log, and assess model behaviors over time to support transparent and verifiable assessments of model behaviors.
    b. **Methods for detecting and attributing LLM outputs:** Developing techniques—such as watermarking or model fingerprinting—to identify and attribute content to its source model provides a foundation for accountability, reduces misinformation risks, and clarifies responsibility.
    c. **Regulatory compliance automation:** Automating processes to ensure that models conform to legal standards, industry guidelines, and ethical principles helps organizations proactively meet accountability requirements while reducing manual oversight burdens.
    d. **Methods for mitigating power concentration in AI:** Investigating mechanisms to prevent the centralization of AI capabilities and influence—such as decentralized governance, open-source contributions, and equitable resource allocation.

14. **Ethics:** Work on AI ethics includes developing methods for integrating ethical considerations into training, evaluation, and decision-making processes, as well as techniques for mitigating harmful outputs and ensuring cultural and long-term ethical consistency.

    a. **Ethics-aware training and fine-tuning:** Research on learning from imperfect ethical datasets, applying ethics-aware data curation methods, and incorporating collective ethical principles into model design.
    b. **Ethical decision-making frameworks:** Developing formal risk-aware, algorithmic harms assessment, and domain-specific ethical decision-making frameworks tailored for large language models and related AI systems.

c. **Mitigating harmful outputs:** Approaches include refining models to reduce the production of dangerous, misleading, or otherwise harmful outputs, employing filtering, red-teaming, and reinforcement learning from human feedback.

d. **Cultural sensitivity and contextual awareness:** Techniques aim to adapt models to diverse cultural contexts and subtle social norms, ensuring that outputs remain appropriate, respectful, and aligned with local values.

e. **Long-term ethical consistency:** Research explores methods for maintaining stable, ethically coherent model behavior over extended periods, including approaches to prevent drift and to preserve core ethical principles despite shifting inputs.

15. **Choosing and operationalizing values in AI:** This area focuses on developing principled methods to identify, justify, and implement value systems within AI models, reconciling diverse ethical priorities, managing conflicts, and creating robust evaluations to ensure models embody chosen values.

    a. **Justify value choices for alignment**: Research includes formulating principled criteria and philosophical foundations that guide why certain values should be encoded into AI systems.

    b. **Manage conflicts between different values:** Efforts here explore approaches like multi-objective optimization or deliberation frameworks to resolve cases where multiple values clash.

    c. **Develop more robust evaluations for which values an LLM encodes:** Researchers design metrics and tests to identify and measure the values present in a model's behavior, outputs, and decision-making processes.

    d. **Pluralistic value alignment:** Strategies focus on simultaneously accommodating multiple, possibly diverse value systems, enabling AI to adapt to different stakeholders or cultural contexts.

    e. **Foundational research on operationalizing values in LLMs:** This includes theoretical and empirical studies on how to incorporate values directly into training procedures, fine-tuning protocols, and model architectures.

16. **Privacy:** This area focuses on identifying and mitigating privacy risks arising from new capabilities and deployment scenarios for LLMs, developing robust conceptual frameworks for privacy definitions, and leveraging AI tools to preserve and enhance privacy in various application domains.

    a. **Identifying emergent privacy risks in new paradigms:** Examining novel attack vectors (e.g., inference time risks) in new paradigms (e.g., retrieval-augmented generation, agent-based interactions, plugin ecosystems) to uncover how these integrations may lead to unexpected disclosures.

    b. **Research on inferring sensitive information from accumulation of innocuous data:** Studying how seemingly harmless data points can be combined to reconstruct sensitive information, enabling adversaries to "weaponize" aggregate inferences against individuals.

    c. **Privacy challenges in complex data scenarios:** Exploring how complex data scenarios, such as cross-lingual and cross-modal transformations (e.g., images, audio, code snippets) can reveal protected content, examining what can be extracted from data presented in alternative formats. This also includes work on context-specific privacy norms, ensuring that privacy measures adapt to different cultural, social, and situational factors rather than relying on one-size-fits-all policies.

    d. **Privacy modeling frameworks:** Developing more precise models of privacy that align with user expectations—potentially informed by human-computer interaction (HCI) research—and grounding these definitions in implementable policies. This also includes formalizing methodologies to characterize and prioritize the worst-case privacy outcomes, moving beyond ad hoc assessments and towards systematic threat modeling frameworks.

    e. **Data encryption tools for model inputs and outputs:** Techniques for encrypting inputs, outputs, and intermediate representations during runtime to ensure confidentiality and prevent unauthorized access to sensitive queries or responses.

17. **Cybersecurity for AI models:** Focuses on protecting model parameters, interfaces, training techniques, and outputs from unauthorized access, extraction, or misuse using cryptographic, architectural, and procedural safeguards. This includes ensuring secure weight storage, hardened access control, oracle protection measures, protecting algorithmic insights, preventing self-exfiltration, and robust data integrity.

    a. **Weight security and key management:** Research focuses on cryptographic techniques for encrypting and securely storing model weights at rest and in transit, hardware-based protections (e.g., trusted execution environments) that ensure the model's parameters cannot be extracted even with physical access, developing systems for isolating

weights behind tightly controlled interfaces, and implementing formal verification of key generation, storage, and rotation protocols.

b. **Access control and interface hardening:** Approaches include creating minimal, verifiable interfaces for weight access, deploying multiparty authorization and cryptographic attestation protocols to guard against model extraction, novel authentication and authorization schemes integrating Zero Trust principles at a granular level (beyond standard IAM tools), and implementing AI firewalls with strict input-output validation.

c. **Model robustness and oracle protection:** Techniques prevent model extraction through inference-only attacks, detect and filter adversarial inputs designed to reconstruct the model or degrade its integrity, and employ adversarial training and advanced input/output "reconstruction" methods to limit the risk that internal representations are inferred from model queries.

d. **Preventing model self-exfiltration:** Methods to ensure that models cannot covertly leak sensitive information about their internal parameters or training data, or copy their own weights to external devices and networks, such as output restrictions, sanitization techniques, or fine-grained monitoring of responses.

e. **Detecting modified models or poisoned data:** Developing methods to detect models that have been maliciously modified or training data which has been poisoned.

f. **Quantifying cyber threats from advanced capabilities:** Threat modeling and evaluation of cyber threats from advanced AI models, whether via autonomy or providing human uplift for more "traditional" cyber capabilities.

18. **Hardware and infrastructure security for AI:** Ensuring the security of AI systems at the hardware and infrastructure level involves protecting model weights, securing deployment environments, maintaining supply chain integrity, and implementing robust monitoring and threat detection mechanisms. Methods include the use of confidential computing, rigorous access controls, specialized hardware protections, and continuous security oversight.

    a. **Confidential computing and environment isolation:** Using trusted execution environments (such as secure enclaves) to ensure that model weights and computations remain confidential and tamper-proof during large-scale AI inference and training. This also involves reducing the attack surface through sandboxed, code-minimal deployments, specialized hardware/firmware stacks, and maintaining verifiable runtime integrity checks.

    b. **Supply chain integrity and secure development:** Ensuring end-to-end verification of hardware and software supply chains through source-verified firmware, SLSA compliance, and secure software development lifecycles tailored for ML-specific infrastructure. This also includes developing automated tooling to continuously verify the provenance and integrity of model components, dependencies, and third-party code used in training and inference pipelines.

    c. **Continuous monitoring, advanced threat detection, and incident response:** Developing ML-driven anomaly detection and logging systems capable of flagging and responding to subtle infiltration attempts or insider threats in real-time. This also includes red-teaming and automated penetration testing frameworks specialized for AI systems, including simulations of zero-day attacks and insider compromises.

    d. **Hardware-integrated monitoring and verification:** Integrating monitoring capabilities directly into hardware, such as secure counters and tamper-evident seals, along with deploying specialized firmware that can detect and respond to attempts at parameter theft or physical attacks. This also includes verification tools, such as hardware-level logging and secured audit trails that remain verifiable under sophisticated tampering attempts for rapid, evidence-based incident response.

    e. **Specialized chips to compute encrypted data:** Designing and deploying hardware accelerators optimized for computations on encrypted data, such as homomorphic encryption schemes, to facilitate efficient encrypted training and inference without exposing plaintext model parameters or sensitive input data outside the protected hardware boundary.

    f. **Tamper-evidence and tamper-proofing:** Implementing tamper-resistant enclosures, seals, and other tamper-evident mechanisms to ensure that any unauthorized physical access or modification attempts are detectable. Such measures help maintain the integrity of hardware components and prevent adversaries from compromising the system at a physical level.

    g. **Datacenter security:** Relevant research focuses on designing and deploying resilient hardware- and software-based defenses to prevent model theft and sabotage. This includes methods like encrypted computation, secure enclaves, continuous anomaly detection, zero-trust architectures, and rigorous supply chain verification to protect against both external intrusions and insider threats.

19. **Improving general understanding of deep learning:** This area focuses on developing rigorous explanations for why deep neural networks learn effectively, uncovering the principles behind generalization, understanding optimization behavior, and analyzing how implicit biases and overparameterization influence performance and safety.
    a. **Theoretical foundations of deep learning:** Constructing mathematical models to explain generalization in deep neural networks despite overparameterization, and studying the influence of network architecture on learning properties.
    b. **Optimization and loss landscape analysis:** Studying the geometry of loss functions and how optimization algorithms navigate them, and examining phenomena such as flat versus sharp minima, and the connection between these properties and robust generalization.
    c. **Implicit bias of optimization algorithms:** Analyzing how algorithms like stochastic gradient descent and related methods influence learned models, and exploring how implicit regularization affects model performance and safety.
    d. **Double descent and overparameterization:** Investigating the double descent risk curve and its implications for model capacity, and how overparameterization can lead to improved generalization.

20. **Research on safety in non-LLM systems:** Exploring safety challenges in non-LLM systems, such as robotics and embodied AI, vision and perception systems, and alternative paradigms for developing artificial intelligence (e.g. whole-brain emulation).
    a. **Safe reinforcement learning for non-LLM systems:** Developing RL algorithms that prioritize safety during exploration and exploitation, with applications in non-LLM systems such as robotics and embedded AI. This includes incorporating safety constraints and risk-sensitive objectives into learning processes.
    b. **Robotics and embodied AI safety:** Designing robust control systems, fail-safe mechanisms, and dependable sensors for physical systems, such as autonomous vehicles, drones, and household robots, to ensure safe human-robot interaction and accident prevention.
    c. **Adversarial robustness in vision and perception systems:** Studying how malicious inputs can deceive image recognition or sensor-based models, and creating defenses—such as adversarial training, certifiable robustness methods, and detection schemes—to maintain reliable perception.
    d. **Whole-brain emulation:** Exploring the theoretical challenges of accurately replicating a human brain's functionality and ensuring that such emulations—if ever realized—adhere to rigorous safety and ethical standards, avoiding unintended cognitive hazards or harmful behavioral patterns.