Sören Mindermann

Curriculum Vitae

Summary

Roles: *Scientific Lead* and later *Advisor to the Chair* of the 30-nation *International AI Safety Report* · Previously *postdoc* at Mila under Yoshua Bengio · PhD in machine learning from the University of Oxford under Yarin Gal.

Publication venues as (equally contributing) first or senior author: *Science* (2x), *Nature Communications*, *PNAS*, *NeurIPS*, *ICML*, *ICLR*, etc. Co-authored work in *Nature*.

Impacts: National-scale policy impacts · minister-level briefings and discussions · Interviews on top-tier TV and news outlets · Invited talks at ML conferences, industry, academia, think tanks and (inter)national government institutions · university-wide research impact awards.

Education

Oct. 2019 – May 2024 **Ph.D. in Machine Learning**, *University of Oxford*.
Supervised by Yarin Gal (Department of Computer Science).

Sept. 2016 – Sept. 2017 **MSc Computational Statistics and Machine Learning**, *University College London*.
- Distinction.
- Thesis under Peter Dayan on hierarchical Bayesian reinforcement learning.

Sept. 2013 – Sept. 2016 **BSc Mathematics**, *University of Amsterdam*, *GPA: 8 (equivalent to **4.0**)*.
- Co-authored a review on efficient Monte Carlo methods in year 2.

Sept. 2012 – Sept. 2016 **BSc Future Planet Studies**, *University of Amsterdam*, *GPA: 7.6 (equivalent to **3.6**)*.
- Natural and social sciences degree. Specialized in economics and governance of resources, energy, water, food, and environment.
- Completed two 3-year degrees simultaneously in 4 years.
- In preparation, self-taught Dutch from no skills to testing at fluent level (C1) in 7 weeks.

Work experience

April 2026 – present **Technical Research Lead**, *Cambridge Programme on AI Science & Policy*, University of Cambridge.

July 2025 – present **Research Affiliate**, *Oxford Martin AI Governance Initiative*.

November 2023 – February 2026 **Postdoctoral Researcher → Scientific Lead → Advisor to the Chair**, *Mila - Quebec AI Institute*.
Postdoc supervised by Yoshua Bengio. Then became the *Scientific Lead*, and later the *Scientific Advisor to the Chair* of the *International AI Safety Report*, a project mandated by 30 nations, the UN, EU and OECD.

July 2019 – October 2019 **AI Governance Fellow**, *University of Oxford*, GovAI - Centre for the Governance of AI.
Wrote economics paper predicting vertical disintegration and AI APIs which now materialized.

July – December 2018 **Research intern**, *University of Toronto*, Vector Institute.
Machine learning for open source game theory under Prof. David Duvenaud and Roger Grosse.

November 2017 – May 2018 **Visiting scholar**, *UC Berkeley*, CHAI group (Russel, Abbeel, Dragan).
Lead author on 'Active Inverse Reward Design'.

October 2017 – November 2017 **Research intern**, *University of Oxford*, Future of Humanity Institute.
Equal 1st author of NeurIPS theory paper on inverse RL with Dr. Stuart Armstrong.

Selected publications

* = equal contribution

∘ = ordered by coin flip

✉ = corresponding author

2025 Y Bengio (Chair), **S Mindermann (Scientific Lead)**✉, D Privitera (Lead Writer), T Besiroglu, R Bommassani, S Casper, Y Choi, +93 authors. *International AI Safety Report 2025*.
CONTRIBUTION: Scientific Lead from the start, covering the 2024 interim, 2025 main, and 2026 editions. Sole corresponding author 2025; last author 2026.

2024 Y Bengio, G Hinton, A Yao, D Song, P Abbeel, Y N Harari, T Darrell, Y Zhang, L Xue, S Shalev-Shwartz, G Hadfield, J Clune, T Maharaj, F Hutter, A G Baydin, S McIlraith, Q Gao, A A, D Krueger, A Dragan, P Torr, S Russell, D Kahneman, J Brauner*✉, **S Mindermann***✉. *Managing Extreme AI Risks Amid Rapid Progress*. In **Science**.
CONTRIBUTION: Performed all conception, writing and organization, together with Jan Brauner.

2023 R Ngo✉, L Chan✉, **S Mindermann**✉. *The Alignment Problem from a Deep Learning Perspective*. In **ICLR**.

2022 **S Mindermann***✉, Muhammed Razzak*, Winnie Xu*, Andreas Kirsch, Mrinank Sharma, Aidan Gomez, Sebastian Farquhar, Jan Brauner, Yarin Gal. *Prioritized Training on Points that are Learnable, Worth Learning, and Not Yet Learnt.* In **ICML.**
CONTRIBUTION: Conceived the algorithm, designed most experiments, led and managed the project.

2021 JM Brauner*✉, **S Mindermann***✉, M Sharma*✉, D Johnston, J Salvatier, T Gavenciak, AB Stephenson, G Leech, G Altman, V Mikulik, AJ Norman, JT Monrad, T Besiroglu, H Ge, MA Hartwick, YW Teh, L Chindelevitch, Gal Y, J Kulveit. *Inferring the effectiveness of government interventions against COVID-19*. In **Science**.
CONTRIBUTION: Led on designing the models and experiments.

Publications as lead author

2025 Y Bengio (Chair), **S Mindermann (Scientific Lead)**✉, D Privitera (Lead Writer), T Besiroglu, R Bommassani, S Casper, Y Choi, +93 authors. *International AI Safety Report.*

2025 EVENT PLANNING COMMITTEE: Y Bengio, T Maharaj, L Ong, S Russell, D Song, M Tegmark, L Xue, Y-Q Zhang.
WRITING GROUP: M Tegmark, **S Mindermann** (one of 3 scientific main authors), S Casper, V Wilfred, W S Lee.
INDIVIDUAL CONTRIBUTORS: +80 others.
The Singapore Consensus on Global AI Safety Research Priorities. Singapore Government.

2024 Y Bengio (Chair), **S Mindermann (Scientific Lead)**, D Privitera (Lead Writer), T Besiroglu, R Bommassani, S Casper, Y Choi, +75 authors. *International Scientific Report on the Safety of Advanced AI: Interim Report.*

2022 **S Mindermann***✉, Muhammed Razzak*, Winnie Xu*, Andreas Kirsch, Mrinank Sharma, Aidan Gomez, Sebastian Farquhar, Jan Brauner, Yarin Gal. *Prioritized Training on Points that are Learnable, Worth Learning, and Not Yet Learnt.* **International Conference on Machine Learning.**

2021 JM Brauner*✉, **S Mindermann***✉, M Sharma*✉, D Johnston, J Salvatier, T Gavenciak, AB Stephenson, G Leech, G Altman, V Mikulik, AJ Norman, JT Monrad, T Besiroglu, H Ge, MA Hartwick, YW Teh, L Chindelevitch, Gal Y, J Kulveit. *Inferring the effectiveness of government interventions against COVID-19*. In **Science**.

2021 **S Mindermann***✉○, Mrinank Sharma*✉○, Charlie Rogers-Smith, Gavin Leech, Benedict Snodin, Janvi Ahuja, Jonas B Sandbrink, Joshua Teperowski Monrad, George Altman, Gurpreet Dhaliwal, Lukas Finnveden, Alexander John Norman, Sebastian B Oehm, Julia Fabienne Sandkühler, Thomas Mellan, Jan Kulveit, Leonid Chindelevitch, Seth Flaxman, Yarin Gal, Swapnil Mishra, Jan Markus Brauner✉, Samir Bhatt✉. *Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe*. In **Nature Communications**.

2021 S Mishra*✉, **S Mindermann***, M Sharma*, C Whittaker*, T Mellan, T Wilton, D Klapsa, R Mate, M Fritzsche, M Zambon, J Ahuja, A Howes, X Miscouridou, G Nason, O Ratmann, G Leech, J Fabienne Sandkuhler, C Rogers-Smith, M Vollmer, H Unwin, Y Gal, M Chand, A Gandy, J Martin, E Volz, N Ferguson, S Bhatt, J Brauner, S Flaxman. *Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England*. In **EClinicalMedicine (The Lancet)**.

2020 Mrinank Sharma*, **S Mindermann***, Jan Brauner*, Gavin Leech, Anna Stephenson, Tomas Gavenciak, Jan Kulveit, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal. *How Robust are the Estimated Effects of Nonpharmaceutical Interventions against COVID-19?* In **NeurIPS (Spotlight talk)**.

2020 A Jesson*, **S Mindermann***, U Shalit, Y Gal. *Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models.* In **NeurIPS**.

2018 **S Mindermann***○ & S Armstrong*○. *Occam's razor is insufficient to infer the preferences of irrational agents.* In **NeurIPS**.

2018 **Mindermann*, S.**, Shah*, R., Gleave, A., Hadfield-Menell, D.. *Active Inverse Reward Design*, AAMAS/ICML workshop on goals in RL.

Publications as senior author

* = equal contribution to senior authorship

2026 Y Bengio (Chair), Stephen Clare, Carina Prunkl, ... +87 authors ... D Privitera, **S Mindermann (Advisor to the Chair)**. *International AI Safety Report 2026*.

2024 Y Bengio, G Hinton, A Yao, D Song, P Abbeel, Y N Harari, T Darrell, Y Zhang, L Xue, S Shalev-Shwartz, G Hadfield, J Clune, T Maharaj, F Hutter, A G Baydin, S McIlraith, Q Gao, A A, D Krueger, A Dragan, P Torr, S Russell, D Kahneman, J Brauner*✉, **S Mindermann***✉. *Managing Extreme AI Risks Amid Rapid Progress.* In **Science**.

2024 E Hubinger*, C Denison*, J Mu*, M Lambert*, M Tong*, M MacDiarmid, T Lanham, D M Ziegler, T Maxwell, N Cheng, A Jermyn, A Askell, A Radhakrishnan, C Anil, D Duvenaud, D Ganguli, F Barez, J Clark, K Ndousse, K Sachan, M Sellitto, M Sharma, N DasSarma, R Grosse, S Kravec, Y Bai, Z Witten
Senior authors block:
M Favaro, J Brauner, H Karnofsky, P Christiano, S R Bowman, L Graham, J Kaplan, **S Mindermann**, R Greenblatt, B Shlegeris, N Schiefer*, E Perez*. *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training.* Anthropic Research.

2023 R Ngo✉, L Chan✉, **S Mindermann**✉. *The Alignment Problem from a Deep Learning Perspective.* In **ICLR**.

2022 G Leech✉, C Rogers-Smith, J Sandbrink, B Snodin, R Zinkov, B Rader, J Brownstein, Y Gal, S Bhatt*, M Sharma*, **S Mindermann***, J Brauner*, L Aitchison*. *Mask wearing in community settings reduces SARS-CoV-2 transmission.* **Proceedings of the National Academy of Sciences (PNAS)**.

2022 G Altman✉, J Ahuja✉, JT Monrad, G Dhaliwal, C Rogers-Smith, G Leech, B Snodin, JB Sandbrink, L Finnveden, AJ Norman, SB Oehm, JF SandkÃŒhler, J Kulveit, S Flaxman, Y Gal, S Mishra, S Bhatt, M Sharma*, **S Mindermann***, J Brauner*. *A dataset of non-pharmaceutical interventions on SARS-CoV-2 in Europe.* **Nature Scientific Data**.

Publications as co-author

2026 A Cloud, M Le, J Chua, J Betley, A Sztyber-Betley, **S Mindermann**, J Hilton, S Marks, O Evans. *Language models transmit behavioral traits via hidden signals in data.* In **Nature** (forthcoming).

2025 A Lynch, B Wright, C Larson, KK Troy, SJ Ritchie, **S Mindermann**, E Perez, E Hubinger. *Agentic Misalignment: How LLMs Could be an Insider Threat.* Anthropic Research.

2025 B Bucknall, S Siddiqui, L Thurnherr, C McGurk, B Harack, A Reuel, P Paskov, C Mahoney, **S Mindermann**, S Singer, V Hiremath, C Segerie, O Delaney, A Abate, F Barez, MK Cohen, P Torr, F HuszÃ¡r, A Calinescu, GD Jones, Y Bengio, R Trager *In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?.* In **ACM - FAccT**.

2025 Y Bengio, M Cohen, D Fornasiere, J Ghosn, P Greiner, M MacDermott, **S Mindermann**, A Oberman, J Richardson, O Richardson, M-A Rondeau, P-L St-Charles, D Williams-King. *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?.* In Arxiv.

2025 F Barez, T Fu, A Prabhu, S Casper, A Sanyal, ABibi, A O'Gara, RKirk, B Bucknall, T Fist, L Ong, P Torr, K-Y Lam, R Trager, D Krueger, **S Mindermann**, J Hernandez-Orallo, M Geva, Y Gal. *Open Problems in Machine Unlearning for AI Safety.* In Arxiv.

2025 J Clymer, I Duan, C Cundy, Y Duan, F Heide, C Lu, **S Mindermann**, C McGurk, X Pan, S Siddiqui, J Wang, M Yang, X Zhan. *Bare Minimum Mitigations for Autonomous AI Development.* In Arxiv.

2024 R Greenblatt*, C Denison*, B Wright*, F Roger*, M MacDiarmid*, S Marks, J Treutlein, T Belonax, J Chen, D Duvenaud, A Khan, J Michael, **S Mindermann**, E Perez, L Petrini, J Uesato, J Kaplan, B Shlegeris, SR Bowman, E Hubinger*. *Alignment faking in large language models.* Anthropic Research.

2023 L Pacchiardi*, A J Chan*, **S Mindermann**, I Moscovitz, A Pan, Y Gal, O Evans, J Brauner*. *How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions.* **ICLR**.

2023 S Kundu, Y Bai, S Kadavath, A Askell, A Callahan, A Chen, A Goldie, A Balwit, A Mirhoseini, B McLean, C Olsson, C Evraets, E Tran-Johnson, E Durmus, E Perez, J Kernion, J Kerr, K Ndousse, K Nguyen, N Elhage, N Cheng, N Schiefer, N DasSarma, O Rausch, R Larson, S Yang, S Kravec, T Telleen-Lawton, T I Liao, T Henighan, T Hume, Z Hatfield-Dodds, **S Mindermann**, N Joseph, S McCandlish, J Kaplan. *Specific Versus General Principles for Constitutional AI.* Anthropic Research.

2023 A Lison, N Banholzer, M Sharma, **S Mindermann**, H Juliette T Unwin, S Mishra, T Stadler, S Bhatt✉, N Ferguson, J Brauner, and W Vach. *Effectiveness assessment of non-pharmaceutical interventions: lessons learned from the COVID-19 pandemic.* In **The Lancet: Public Health**.

2021 A Jesson✉, **S Mindermann**, Y Gal, U Shalit. *Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding.* In **International Conference on Machine Learning**.

2021 Gideon Meyerowitz-Katz✉, Samir Bhatt, Oliver Ratmann, Jan Markus Brauner, Seth Flaxman, Swapnil Mishra, Mrinank Sharma, **S Mindermann**, Valerie Bradley, Michaela Vollmer, Lea Merone, Gavin Yamey. *Is the cure really worse than the disease? The health impacts of lockdowns during COVID-19.* In **BMJ Global**.

2021 Tomas Gavenciak*, Joshua Teperowski Monrad*✉, Gavin Leech, Mrinank Sharma, **S Mindermann**, Jan Markus Brauner, Samir Bhatt, Jan Kulveit*. *Seasonal variation in SARS-CoV-2 transmission in temperate climates.* In **PLOS Computational Biology**.

Awards

2022 **MPLS Impact Award.** Awarded for impact of research on "Understanding effectiveness of interventions against Covid-19 using Bayesian models". 1st prize among Oxford researchers in any career stage across all MPLS departments (STEM and life sciences). The award forms a basis for the REF studies and has been given only 28 times historically, including for the largest contribution to creating the *R* programming language, and to Nobel laureate Roger Penrose.

2021 Edward Chapman Research Prize 2021. Best first-author paper in the natural sciences at Magdalen College, Oxford.

2019 DeepMind-Oxford Scholarship (£95,885 for 4 years).

2019 Long-Term Future Fund grant (£60,000 for 4 years)

2016 Pareto Fellowship.

2014 Heinrich Böll Foundation scholarship for academic and social achievements. (€36,000)

2011 German Mathematical Society Abitur Prize, 1st among 142 students.

Policy impact

2025 The International AI Safety Report 2025 is acknowledged as the primary evidence base for the *California Report on Frontier AI Policy* commissioned by Governor Newsom. The report directly led to SB-53, the first enacted US law targeting frontier AI.

2025 The International AI Safety Report 2025, for which I am the Scientific Lead, was presented at the international AI Action Summit hosted by the French government. Numerous policymakers were briefed on the report.

2025 The **UK's new minister of Science**, Innovation and Technology (Peter Kyle) cited the International AI Safety Report (2025) in his speech at the Munich Security Conference.

2024 The **UK's minister of Science**, Innovation and Technology presented the international AI safety report to ministers of other countries and to other high-level representatives at the Seoul AI Summit. I am the scientific lead of this report.

2023 *Managing Extreme AI Risks Amid Rapid Progress* which I co-led, briefed to **Germany's minister of health**.

2021 I presented statistical modeling work on mask-wearing at the UK Cabinet Office to support the UK's plan for fall 2021. (I co-supervised this paper.)

2021 Preprint cited in the **German federal bill** that decided the national lockdown in force as of May 2021. (One of three papers cited.)

2021 Some COVID-19 papers on which I was (equally contributing) first author have been presented at the **WHO**, the modeling groups of the **Africa CDC** and the UK's Scientific Advisory Group for Emergencies (**SAGE**), and the **House of Representatives** of the Netherlands.

2020 I presented our work on interventions against COVID-19 transmission to the modelling group of the Africa CDC.

Interviews given on TV and newspapers

2025 *CBS - 60 Minutes.* Interview on background covering my evaluations research on alignment and methodological challenges.

2025 *Transformer News.* Discussed my work on agentic misalignment and methodological challenges.

2025 *Transformer News.* Discussed my work on alignment faking.

2024 *Spiegel (Kindermagazin).* 'How alignment works'.

2021 *Monitor TV magazine on **ARD*** (German equiv. of BBC, ca. 3M viewers per episode). Talked about preprint covering COVID's 2nd wave.

2021 ***ITV Peston*** (the flagship political TV program of ITV): Talked about paper covering COVID's 2nd wave.

2021 ***DW News*** (TV). Covid interventions in the 2nd wave.

2021 *Süddeutsche Zeitung (top-tier German newspaper).* Interview about government interventions in COVID's second wave.

2023 *Analytics India Magazine.* Discussed AI risk management.

2021 *NRC Handelsblad.* Talked about paper on government interventions in COVID's first wave.

2023 *Prioritäten* (podcast). Sören Mindermann on the problem of aligning AI (translated).

2021 *Alan Turing Institute* (podcast): Government interventions in COVID's first wave.

Invited talks

2026 **International Association on the Safety and Ethics of AI**, *Panel: Open Weight Model Safety.*

2026 **Vector Institute (University of Toronto)**, *The International AI Safety Report 2026.*

2026 **FAR AI - Alignment Workshop**, *The International AI Safety Report 2026.*

2025 **Simon Institute for Long-Term Governance**, *Keynote: International AI Safety Report (presented to Geneva UN diplomatic missions staff).*

2025 **Turkey Ministry of Industry and Technology and Ministry of Foreign Affairs**, *Led briefing on the International AI Safety Report to policymakers at the Director General level.*

2025 **Tarbell Fellowship; London Initiative for Safe AI**, *International AI Safety Report.*

2025 **Vector Institute (University of Toronto)**, *The International AI Safety Report.*

2024 **IBM Research**, *The Alignment Problem from a Deep Learning Perspective.*

2023 **London Initiative for AI Safety**, *Managing AI Risks in an Era of Rapid Progress.*

2023 **Oxford University - Philip Torr Vision Group**, *The Alignment Problem from a Deep Learning Perspective.*

2023 **Future of Life Institute**, *The Alignment Problem from a Deep Learning Perspective.*

2023 **University of Amsterdam**, *The Alignment Problem from a Deep Learning Perspective.*

2023 **University of Edinburgh**, *The Alignment Problem from a Deep Learning Perspective.*

2023 **UC Berkeley (CHAI)**, *The Alignment Problem from a Deep Learning Perspective.*

2022 **Meta AI**, *Prioritized training on points that are learnable, worth learning, and not yet learned.*

2022 **USC (Cutelab)**, *Prioritized training on points that are learnable, worth learning, and not yet learned.*

2022 **University of Oxford—AI for Agent-Based Modeling seminar**, *Inferring the Effectiveness of government interventions against COVID-19.*

2021 **Cohere.ai**, *Prioritized training on points that are learnable, worth learning, and not yet learned.*

2021 **ETH Zürich**, *Government interventions in the second wave.*

2021 **Imperial College - MRC Centre for Global Infectious Disease Analysis**, *Government interventions in the second wave.*

2020 **Africa CDC**, *Inferring the effectiveness of government interventions against COVID-19.*

2020 **German Centre for Infection Research**, *Inferring the effectiveness of government interventions against COVID-19.*

2020 **NeurIPS COVID Symposium Spotlight**, *How Robust are the Estimated Effects of Nonpharmaceutical Interventions against COVID-19?.*

Peer reviewing

2026 Reviewer, Nature

2026 Reviewer, Nature

2025 Area Chair, ICML workshop on Technical AI Governance

2023 Reviewer, ICML

2022 Reviewer, ICLR

2022 Reviewer, NeurIPS

2022 Reviewer, ICML

2021 Reviewer, ICLR

2021 Reviewer, NeurIPS

2021 Reviewer, ICML

2020 Reviewer, Nature Machine Intelligence

2020 Reviewer, NeurIPS

2020 Reviewer, ICML

2018 Reviewer, NeurIPS Smooth Games Optimization and Machine Learning Workshop.

Volunteer service

2014—present **Wikipedia contributor**.
Authored articles such as *AI Alignment* to improve accessible education.

April 2013– April 2014 **Volunteer**, *New Harvest*.
Built a database and map of scientists and identified grant-makers for alternative protein research.

2012–2015 **Board member / later treasurer**, *Interdisciplinary student association (Spectrum)*.
Organized student trips.