Charbel-Raphaël Segerie

Paris, France

charbel@cesia.org | +33 6 95 69 28 86 | linkedin.com/in/charbel-raphael-segerie

Policy leader, AI researcher, and institution-builder working at the intersection of technical AI research, international governance, and education. Executive Director of CeSIA, France's leading AI safety organization. OECD AI Expert. Official evaluator for the EU AI Act Code of Practice on harmful manipulation risks. Initiator of the Global Call for AI Red Lines—endorsed by 12 Nobel laureates and presented at both the UN General Assembly and UN Security Council. Creator of the first EU university-accredited course on general-purpose AI safety. Founder of ML4Good, an EU Commission-funded programme upskilling professionals and researchers on technical AI safety and AI governance, replicated 20+ times worldwide.

Current Positions

Executive Director & Co-Founder 2024 – present
Centre pour la Sécurité de l'IA (CeSIA), Paris

- Leading a 7-person team across three strategic tracks: France (domestic AI governance and G7 influence), Europe (EU AI Act implementation), and International (Red Lines initiative for binding international agreements).
- Scientific Director and co-author of the AI Safety Atlas, used by 1,000+ students worldwide.
- Contributor to all three rounds of the EU AI Act's Code of Practice; recommendations included verbatim in the final draft.
- Published on the OECD AI Policy Observatory with Stuart Russell on global AI red lines.

Official Evaluator, EU AI Office GPAI Code of Practice 2026 – 2028
Lot 4: Harmful Manipulation Risks — consortium with Apart Research, Transluce & EquiStamp. Focus on risk modeling.

OECD AI Expert 2024 – present
ONE AI Expert Network, Organisation for Economic Co-operation and Development

Head Teacher, Turing Seminar on AGI Safety 2022 – present
ENS Paris-Saclay (MVA Master) & ENS Ulm

- Created the first university-accredited course on general-purpose AI safety in the EU, at a time when virtually no European university offered such training. Course available on YouTube; textbook: the AI Safety Atlas.

Founder & Curriculum Designer, ML4Good 2022 – present
EU Commission-funded programme upskilling professionals and researchers on technical AI safety and AI governance

- Designed the curriculum and personally led the initial bootcamps in France. The programme has since been replicated 20+ times across Europe, Latin America, and beyond, with a 98% participant recommendation rate. Hundreds of professionals have transitioned to AI safety careers through CeSIA and ML4Good. Alumni now hold positions at the EU AI Office, LawZero (Y. Bengio), Mistral, MATS Research, GPAI Policy Lab, and Seldon Lab.

Selected Achievements & Impact

- **Global Call for AI Red Lines** (initiator & co-lead): 12 Nobel laureates, 10 former heads of state, 200+ total signatories; presented at the UN General Assembly by Nobel Peace laureate Maria Ressa and highlighted by Yoshua Bengio at the UN Security Council; 300+ media mentions (NYT, BBC, Le Monde, TIME, NBC, AP, AFP).
- **India AI Impact Summit** (Delhi, Feb. 2026): Convened workshop "Defining and Governing Unacceptable AI Risks: Toward Global Convergence on AI Red Lines" with officials from the EU, Japan, Singapore, Brazil, Denmark, Canada, and UNESCO.
- **AI Action Summit** (Paris, Feb. 2025): Co-organized AI Safety Symposium with keynotes by Y. Bengio and S. Russell; hosted official side-event with GovAI, METR, and the UN; CeSIA cited 11 times in the final consultation report. Contributed to the 2nd India–France AI Policy Roundtable (Sciences Po).
- **IASEAI Workshop at UNESCO Paris** (Feb. 2026): Reached consensus on political bottlenecks for AI red lines with international participants and OECD representatives.
- **Frontier Safety Frameworks Coordination Workshop**: Co-organized with FAR.AI at AW London (Mar. 2026) with representatives from Anthropic, DeepMind, and other frontier labs, comparing risk thresholds and discussing harmonization.
- **Presentations at the French Senate** (Palais du Luxembourg, twice), including in front of the French AI Minister, on AI risks and governance.

- **Media and public outreach**: CeSIA featured or cited in Le Monde, NBC, The Verge, Les Échos. Appeared on TF1 (France's main TV channel, Innovation Days), France Inter (France's main public radio), and numerous podcasts and YouTube programmes. Collaborated on a YouTube video on AI risks reaching 4 million views.

Publications

Martinet, Abecassis, **Segerie**, Bengio et al. (2025). "A Blueprint for Multinational Advanced AI Development." *CeSIA & Oxford Martin AIGI Report.*

Bucknall, **Segerie**, Bengio et al. (2025). "In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?" *ACM FAccT 2025.*

Grey & **Segerie** (2025). "The AI Risk Spectrum: From Dangerous Capabilities to Existential Threats." *arXiv:2508.13700.*

Grey & **Segerie** (2025). "Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods." *arXiv:2505.05541.*

Mariaccia, **Segerie** & Dorn (2025). "The Bitter Lesson of Misuse Detection." *arXiv:2507.06282.*

Dorn, Variengien, **Segerie** & Corruble (2024). "BELLS: Benchmarks for the Evaluation of LLM Safeguards." *NextGenAISafety @ ICML 2024.* Included in the OECD catalogue of tools for trustworthy AI.

Casper, **Segerie** et al. (2023). "Open Problems and Fundamental Limitations of RLHF." *Transactions on Machine Learning Research (TMLR).*

Segerie & Gédéon (2024). "Constructability: Plainly-Coded AGIs May Be Feasible in the Near Future." *CeSIA Technical Report.*

Laurençon, **Ségerie**, Lussange & Gutkin (2024). "Continuous Time Continuous Space Homeostatic RL." *ENS / BITS Pilani collaboration.*

Laurençon, **Ségerie**, Lussange & Gutkin (2021). "Continuous Homeostatic RL for Self-Regulated Agents." *arXiv:2109.06580.*

Grey, **Segerie** et al. (2025). *AI Safety Atlas.* CeSIA. ai-safety-atlas.com.

Education

MVA Master (Mathematics, Vision, Learning) 2020–2021
ENS Paris-Saclay

Engineering Degree 2017–2019
École Nationale des Ponts et Chaussées

Previous Experience

Head of the AI Safety Unit, EffiSciences, Paris 2022–2024
Led the AI safety division. Delivered courses at ENS Paris-Saclay and ENS Ulm. Organized hackathons and research mentoring at Collège de France, École 42, Meta, and ENS Ulm. Supervised student research projects on interpretability and AI safety.

CTO, Omniscience (startup), Paris 2021–2022
Built a search engine for semi-automated literature reviews using NLP and semantic retrieval.

Research Intern, Inria Parietal & NeuroSpin (CEA), Saclay 2021
Machine learning for neuroimaging: EEG/fMRI statistical methods and brain–computer interfaces.