Diogo Cruz

Curriculum Vitae

Rua do Vale nº9
2400-768 Amor
Portugal
☎ *(+351) 963 682 728*
✉ *diogo.abc.cruz@gmail.com*
🖱 *www.diogo-cruz.github.io*

General Education

2020-2025 **Physics PhD (Quantum Computing)**, *Instituto Superior Técnico (IST)*, Portugal.
Developed quantum algorithms to solve nonlinear partial differential equations; quantum error correction; and tackle problems for near-future quantum devices.

2025 **Visiting Scholar**, *UC Berkeley*, USA.

2024 **Visiting Student**, *MIT*, USA.

2019-2020 **Graduate Research Assistant**, *Instituto de Telecomunicações*, Portugal.

2014-2019 **Integrated Master Degree (MSc) in Engineering Physics (MEFT)**, *Instituto Superior Técnico*, Lisboa.
Bachelor's and Master's GPA of 18.4 and 18.1/20, resp.; **one of the top students in the course**.
- **Academic Excellence Diploma** in 2014/2015 and 2015/2016;
- **Academic Merit Diploma** in 2016/2017 and 2018/2019;
- **1st Honorable Mention of "Academic Excellence in MEFT"** 2016/2017.

2018 **Erasmus at École Polytechnique Fédérale de Lausanne (EPFL)**, Switzerland.

AI Safety

2025 **Algoverse AI Safety Fellowship**, *Mentor*.
Mentoring 2 ongoing projects, each aiming for a workshop paper, totaling 6 mentees:
1. How long contexts lead LLM agents to have different capability and safety behaviors;
2. Operationalizing situational awareness in LLM agents, and building a benchmark for it.

2025 **SPAR Fall 2025 - AI Safety Research Program**, *Mentor*.
Mentoring 3 ongoing projects, each aiming for a workshop paper, totaling 10 mentees:
1. Developing methods to measure how LLM agents gradually abandon or modify their original goals during extended tasks over long interactions;
2. Investigating how LLM agents behave when their tools fail or become unreliable, and whether this triggers deception, reward hacking, or unauthorized actions;
3. Testing whether fine-tuning models on political text causes emergent misalignment.

2025 **Introduction to AI Evaluations**, *Invited Speaker*, EAGxSãoPaulo, Brazil.

2025 **CHAI Internship**, Berkeley.
Developed methods to train more robust and performant probes when the amount of labeled samples is limited (as is the case for hard-to-classify or superhuman tasks). Preparing conference paper.

2025 **UK AISI bounty programme**, *Contractor*.
Implemented better scaffolding to study agents solving complex tasks.

2024-2025 **Catalyze AI Safety Incubation Program**, *Phases 1 & 2*, London.
Explored evaluation approaches for autonomous AI agents.

2024-2025 **SPAR Spring 2025: Prompt Attacks in Unlearning Methods Project**, *Mentor*.
3 mentees, resulted in COLM SoLaR Workshop paper, `https://arxiv.org/abs/2506.10236`.

2024-2025 **AI Safety Camp 10: Multi-turn Jailbreaks Project**, *Research Lead*.
5 mentees, resulted in COLM SoLaR Workshop paper, `https://arxiv.org/abs/2508.07646`.

2024 **Research Engineers Club**, *Safe AI London*.
Replicated WMDP benchmark paper.

2024 **Pivotal Research Fellowship**, *Pivotal*.
Resulted in TMLR paper. `https://arxiv.org/abs/2505.21552`.

2024 **AI Safety Fundamentals: Governance course**, *Participant & Facilitator*, Bluedot Impact.

2023, 2024 **AI Safety Fundamentals: Alignment course**, *Participant & Facilitator*, Bluedot Impact.

2024 **AI Safety, Ethics, and Society**, *Facilitator*, CAIS.

2023 **ML Safety Scholar Programs**, *CAIS*.

2023 **AI Safety Hub Labs**, *Team Leader*, Oxford.
Neurips SoLaR Workshop paper, `https://arxiv.org/abs/2311.04046`.

2017 **Learning from Data**, *Caltech*, virtual.

2017 **CS231n: Convolutional Neural Networks for Visual Recognition**, *Stanford*, virtual.

Teaching

2022-2023 **Techniques of Mathematical Physics**, *Physics Engineering*, IST.

2022-2023 **Quantum Mechanics**, *Aerospace, Naval, Mechanical Engineering*, IST.

2017-2018 **Electromagnetism and Optics**, *Aerospace, Naval Engineering*, IST.

Selected Publications

See *Google Scholar* for more.

AI Safety

2023 **Reinforcement Learning Fine-tuning of Language Models is Biased Towards More Extractable Features**, *Cruz, D. et al*, arXiv: 2311.04046, 2023.

2024 **Understanding the learned look-ahead behavior of chess neural networks**, *Cruz, D.*, arXiv: 2505.21552, TMLR, 2024.

2025 **Prompt Attacks Reveal Superficial Knowledge Removal in Unlearning Methods**, *Jang, Y. et al*, arXiv: 2506.10236, 2025.

2025 **Multi-Turn Jailbreaks Are Simpler Than They Seem**, *Yang, X. et al*, arXiv: 2508.07646, 2025.

Quantum Computing

2019 **Efficient quantum algorithms for GHZ and W states, and implementation on the IBM quantum computer**, *Cruz, D. et al*, Advanced Quantum Technologies 0 (0), 1900015, 2019.

2023 **Quantum Error Correction via Noise Guessing Decoding**, *Cruz, D.; Monteiro, F. A.; Coutinho, B. C.*, IEEE Access, vol. 11, pp. 119446-119461, 2023.

2023 **Superresolution of Green's functions on noisy quantum computers**, *Cruz, D.; Magano, D.*, Phys. Rev. A 108, 012618, 2023.

2023 **A Living Review of Quantum Computing for Plasma Physics**, *Amaro, O.; Cruz, D.*, arXiv: 2302.00001, 2023.

Selected Awards

2022 **$3000 1st Prize in Classiq Coding Competition's Hamiltonian Exponentiation Challenge**.

2022 **Winner of Hackathon QCHACK 2022 (QuTech Challenge)**, *Stanford*, United States.

2021 **Winner of Hackathon iQuHACK 2021 (D-Wave Challenge)**, *MIT*, United States.

2019 **Grant from Gulbenkian Program "New Talents in Quantum Technologies"**.

2018 **2nd place for "IBM Q Best Paper Award" (DOI: 10.1002/qute.201900015.)**.

2013-2014 **Astronomy Olympiad**.
- **1st place** at the national level;
- **Honorable Mention** in the IOAA in 2014 (international) - best result ever for Portugal.

2013-2014 **Physics Olympiad**.
- **Honorable Mention (Top 10)** at the national level;
- **Honorable Mention** in the XLV IPhO in 2014, in Kazakhstan (international).