Lize Alberts (she/her)

Personal Website lize.alberts@cs.ox.ac.uk

DPhil Candidate in Computer Science at the **University of Oxford**, specialising in human-AI collaboration optimisation. Student Researcher at **Google**, focusing on context-dependent large language model (LLM) alignment and evaluation. **Research Fellow** in AI Ethics at Stellenbosch University. **Multidisciplinary expert** bridging computer science, HCI, machine learning, cognitive science, computational linguistics, philosophy of mind, the social sciences, and human-centred design to address complex AI alignment challenges and contribute to the broader field of AI safety.

Publications

- Manzini, A., Gabriel, I., Ringel Morris, M., Alberts, L., Keeling, G., Vallor, S. (2024) "The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships", *Accepted at the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Alberts, L., Lyngs, U., Van Kleek, M. (2024). "Computers as Bad Social Actors: Dark and Anti-Patterns in Interfaces that Act Socially", *Proceedings of the ACM on Human-Computer Interaction, Volume 8, Issue CSCW1, 202, pp 1–25*. DOI
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L.A., Reiser, V., Iqbal, H., Tomasev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., ... Manyika, J. 2024. "The Ethics of Advanced AI Assistants". *Google DeepMind*. DOI
- Lize Alberts, Geoff Keeling, Amanda McCroskery (2024). "Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect", Under review at *AAAI Conference on AI, Ethics, and Society (AIES)*. arXiv:2401.09082v2. DOI
- Lyngs, U., Lukoff, K., Slovak, P., Inzlicht, M., Freed, M., Andrews, H., Tinsman, C., Csuka, L., Alberts, L., ... Shadbolt, N. (2024). "I finally felt I had the tools to control these urges": Empowering Students to Achieve Their Device Use Goals With the Reduce Digital Distraction Workshop. *Proceedings of the CHI Conference on Human Factors in Computing Systems, 251, pp 1–23*.DOI
- Alberts, L., Lyngs, U., & Lukoff, K. 2024. "Designing for Sustained Motivation: A Review of Self-Determination Theory in Behaviour Change Technologies", accepted at *Interacting with Computers*. arXiv.2402.00121. DOI
- Lukoff, K., Lyngs, U., & Alberts, L. 2022. "Designing to Support Autonomy and Reduce Psychological Reactance in Digital Self-Control Tools", in *Self-Determination Theory in HCI: Shaping a Research Agenda. Workshop at the ACM CHI Conference on Human Factors in Computing Systems (CHI'22)*. PDF
- Alberts, L. 2020. "Not Cheating on the Turing Test: Towards Grounded Language Learning in Artificial Intelligence", master's thesis. arXiv.2206.14672. DOI
- Alberts, L. 2019. "Meeting Them Halfway: Altering Language Conventions to Facilitate Human-Robot Interaction". *Stellenbosch Papers in Linguistics Plus*, 56:97-122. DOI

Work Experience

Student Researcher July-October 2023

Google London, UK

I was headhunted to work as a Student Researcher at Google. I led a project anticipating potential context-specific harms arising from human-agent interaction. I developed a novel evaluation framework and benchmark to assess models' ability to behave in contextually appropriate ways.

Leader of Royal Institution Masterclass in Computer Science 2023-
The Royal Institution, UK
I have designed a Masterclass on *Dark Patterns* and behavioural design ethics that I lead annually for the Royal Institution.

Graduate Research Assistant 2020-2021
Responsible Technology Institute, University of Oxford, UK
I worked as an RA for the EPSRC-funded RoboTIPS project. Using Responsible Innovation (RI) methodology, I assisted with **multi-stakeholder engagement** and constructing **possible scenarios for future technologies**.

Lecturer and Teaching Assistant

- *University of Oxford, UK* 2020-2021
  Software Engineering: *Interaction Design*, *Requirements Engineering*
  Comp. Science: *Ethical Computing in Practice*, *Ethics & Responsible Innovation*
- *Stellenbosch University, ZA* 2018-2019
  I helped administer two Philosophy courses and gave seminars thrice a week for groups of 5 to 80 first/third-years.
- *North-West University, ZA* 2015-2017
  I was course assistant for several Humanities courses and led weekly seminars.

Education

D.Phil. in Computer Science Finalising—Passed Confirmation Viva
University of Oxford, UK
My DPhil research focuses on optimising **user-AI collaboration**, with an emphasis on **LLM-based agents** and socially interactive interfaces. My contributions include defining conditions for respectful and constructive human-AI interaction, supporting individual autonomy, creativity, and wellbeing, and design personalised (context-aware) **alignment techniques**. A key aspect is anticipating and mitigating potential risks in everyday human-agent interactions. By bridging theoretical frameworks with practical applications, this work contributes to the broader field of AI safety and alignment.

M.A. (Thesis) in Philosophy Dec 2020
Stellenbosch University, ZA
Type: Full research thesis (140 pages)
Domains: **Philosophy of Mind and Language**, **Cognitive Science**, Cognitive and Computational Linguistics, **NLP/NLU**
Title: 'Towards grounded language learning in artificial intelligence'
Final grade: **Distinction 84%**

B.A. (Hons) in Philosophy Mar 2019
Stellenbosch University, ZA and *University of Bristol, UK (Study Abroad)*
Bristol: Philosophical Issues of the Physical Sciences, Probability and Rationality
Stellenbosch: Nietzsche, Philosophy of Language Average: **80%**
Dissertation: 'Altering language conventions to facilitate **human-robot interaction'** (18566 words) Mark: **88%**
Final grade: **Cum Laude 82%**

B.A. in Humanities Mar 2018
North-West University (Potchefstroom), ZA
Majors: **Philosophy**, History of Art, **Social Anthropology**
Minors: English, History, Communication Studies
Extra modules: 9 (152 extra credits)
Final grade: **Distinction 82% (Graduated top of class)**

Eldoraigne High School, *ZA*
Final marks: IT **(Object-Oriented Programming, SQL) (87%)**, **Maths (83%)**, English (85%), Afrikaans (89%), Life Orientation (81%), **Design (97%)**, Dramatic Arts (81%),
Nov 2012 Final grade: **Distinction 87%**

Talks and Panels

- Invited panellist on **Dark Patterns as Hostile Scaffolding** 28 June 2024
  Scaffolding Bad Working Group, University of KwaZulu-Natal, UK
- Invited panellist on **Balancing the benefits and harms of AI** 25 May 2024
  Kellogg College, University of Oxford, UK
- **Computers as Bad Social Actors** 8 Mar 2024
  Princeton ETHICOM AI & Ethics Seminar, Princeton University, NJ, USA
- **Dark and Anti-Patterns in Social Interfaces** 13 Sep 2023
  Google-Cambridge University Reading Group on Conversational Agents, Google London, UK
- **Dark and Anti-Patterns in Social Interfaces** 10 Sep 2023
  TL/UXR, Google, London, UK
- **Respect as a Value for Human-Computer Interaction** 21 Mar 2022
  Institute for Ethics in AI gathering with Accenture, University of Oxford, UK
- **Meeting Them Halfway: Altering Language Conventions for Human-Computer Interaction** 28 Oct 2019
  Leverhulme Centre for the Future of Intelligence, Cambridge University, UK

Research Workshops

Google's Techno-Moral Scenarios for Responsible Innovation 17-19 Oct 2023
Google Research/Responsible Innovation, Google Amsterdam, the Netherlands
I was invited to join a select panel of experts from academia and industry to exchange views on using responsible innovation to advance a vision of AI that is empowering, meaningful, and beneficial. We practised using a combination of RI methods (e.g., design fiction, speculative design, and Google's own *techno-moral scenario* approach) to imaginatively and reflexively engage with ethical concerns regarding new technologies.

Generative AI and Its Impact on Human Creativity 1 Feb 2024
Google and the Graziano Lab, Princeton University, NJ, US
I was invited to join a select panel of scholars and Googlers to produce a set of responsible and ethical recommendations regarding Gen(erative) AI. We reflected on the nature and processes of human creativity, and how GenAI may affect, and ideally support it.

Self-Determination Theory in HCI: Shaping a Research Agenda 9 May 2022
The ACM CHI Conference on Human Factors in Computing Systems, Hawaii, US
I participated in a CHI workshop with other HCI researchers to explore how self-determination theory may aid the design of user-centred technologies.

Skills

- **Programming:** Python (for deep learning, data science); MATLAB (for ML); R; Octave; Object Pascal; C; C#; HTML; JavaScript; SQL; PHP
- **GenAI-specific research**: RLHF approaches and self-correction strategies for LLMs; prompt engineering, fine-tuning, alignment; creating novel benchmarks for LLM eval; machine learning/deep learning theory and practice; using responsible innovation methods to evaluate ethical concerns surrounding foundation models.

- **Interface Design:** Interaction design, requirements engineering, user-centred design, value-sensitive design, speculative design, graphic design
- **Empirical research:** Quantitative and qualitative social scientific research methods; thematic analysis; systematic and scoping reviews
- **Applied research:** Cross-disciplinary research; critical and conceptual analysis; applied ethics; evaluating the potential risks/impacts of AI technologies in multidisciplinary teams, multi-stakeholder engagement
- **Academic and Creative Writing:** Writing/reviewing academic papers, poetry

---

Additional Certification

- **Generative AI with LLMs** by deeplearning.ai on Coursera — Jun 2024
- **Neural Networks and Deep Learning** by deeplearning.ai on Coursera
  Grade: **100%** — Dec 2018
- **Diploma in Web Development** by Shaw Academy — Jul 2016

Organisational Experience

Leader of Inter-Departmental Reading Group on LLMs 2023-
The University of Oxford, UK
I lead a biweekly reading group on emerging research on LLMs where Oxford University researchers from Computer Science, Engineering and the Oxford Internet Institute (OII) share their work, discuss new developments in the field, and take GenAI courses.

Co-founder and Organiser of RTI Student Network 2020-
The Responsible Technology Institute (RTI), The University of Oxford, UK
The RTI is an international centre of excellence on responsible technology. I co-founded its international student network connecting researchers interested in RI. I help to organise reading groups, expert panel discussions, and work-in-progress seminars.

President of Oxford University CompSoc 2021-2022
The Oxford University Computer Society (CompSoc)
I was the first female president of CompSoc (est. 1978), one of Oxford University's largest and oldest societies for computer science-related interests. I delegated work, engaged with industry sponsors, and helped organise a university-wide hackathon.

Awards and Achievements

Graduate Lighthouse Scholarship 2020
The RTI and Department of Computer Science, University of Oxford, UK
I received this highly competitive scholarship from Oxford University to fund my DPhil.

Awards for Academic Achievement

- *Department of Philosophy, North-West University, ZA* — Mar 2018
  Merit certificate for excellent academic achievement in Philosophy (top of class)
- *School of Social and Gov. Studies, North-West University, ZA* — Mar 2016
  Merit certificate for best academic achievement in Medical Anthropology
- *School of Social and Gov. Studies, North-West University, ZA* — Mar 2016
  Merit certificate for best academic achievement in Urban Anthropology

Member of Golden Key International Honour Society 2016
North-West University (Potchefstroom), ZA
I was invited for being ranked in the top 15% of North-West University's sophomores.
