Elena Kowalski

AI Ethics & Evaluation Specialist | LLM Alignment Engineer
Berlin, Germany | +49 177 555 8942 | elena.kowalski@example.com
linkedin.com/in/elenakowalski-ai | github.com/ekowalski-alignment

---

Professional Summary

Thoughtful AI Ethics & Evaluation Specialist with 6+ years of experience in responsible AI development, focusing on evaluation methodologies and alignment techniques for large language models. Expertise in developing frameworks for assessing model capabilities, detecting potential harms, and ensuring AI systems align with human values and intentions.

Professional Experience

Lead AI Alignment Engineer
ResponsibleAI GmbH, Berlin, Germany (Jan 2021 - Present)
- Developed comprehensive evaluation framework for assessing LLM capabilities across 50+ dimensions of performance and safety. - Created novel RLHF pipelines improving model alignment with human preferences by 40%. - Designed red-teaming methodologies uncovering 30% more potential vulnerabilities than standard approaches. - Led cross-functional team of 6 researchers focusing on alignment techniques and evaluation methodologies.

AI Ethics Researcher
European AI Institute, Munich, Germany (Mar 2018 - Dec 2020)
- Conducted research on bias detection and mitigation in language models. - Developed benchmarks for evaluating fairness across demographic groups. - Created guidelines for responsible deployment of AI systems in high-stakes domains. - Collaborated with policy experts on developing technical standards for AI governance.

Machine Learning Engineer
TechForward, Vienna, Austria (Aug 2016 - Feb 2018)
- Implemented NLP systems for content moderation and toxicity detection. - Developed evaluation frameworks for measuring model performance across languages. - Created data collection methodologies ensuring diverse representation in training data.

Technical Skills

- **Evaluation Approaches:** Benchmark Design, Red Teaming, Adversarial Testing, Preference Learning
- **Alignment Techniques:** RLHF, Constitutional AI, DPO, SFT
- **Bias & Fairness:** Bias Metrics, Fairness Constraints, Counterfactual Testing
- **Programming Languages:** Python, R, SQL
- **ML Frameworks:** PyTorch, TensorFlow, HuggingFace, ML Fairness
- **Analysis Tools:** Statistical Testing, Uncertainty Quantification, Interpretability Methods
- **Human Feedback Systems:** Preference Collection, Annotation Platforms

---

Education

Doctor of Philosophy (PhD), AI Ethics

Technical University of Munich, Germany (Graduated: Dec 2017)

- Dissertation: “Evaluation Frameworks for Responsible Large Language Models” - Awarded with summa cum laude distinction

Master of Science, Cognitive Science
University College London, UK (Graduated: Jun 2014)
- Thesis: “Cognitive Biases in Human-AI Interaction” - Distinction Award recipient

Bachelor of Arts, Philosophy and Computer Science
Humboldt University of Berlin, Germany (Graduated: May 2012)
- Double major with focus on ethics of technology - Graduated with honors

---

Research Publications

- Kowalski, E., et al. (2023). “Comprehensive Evaluation Methods for Large Language Models.” *FAccT Conference.*
- Kowalski, E., et al. (2022). “Aligning Language Models with Human Preferences: A Survey.” *NeurIPS.*
- Kowalski, E., et al. (2021). “Detecting and Mitigating Bias in LLMs.” *ACL.*
- Kowalski, E., et al. (2020). “Evaluation Frameworks for Responsible AI.” *AIES Conference.*

Professional Service

- Ethics Committee, European Association for AI (2022-present)
- Program Committee: FAccT, AIES, NeurIPS Ethics Track (2019-present)
- Co-organizer, Responsible AI Workshop at NeurIPS 2022
- Reviewer, Journal of AI Research Ethics

Invited Talks & Presentations

- Keynote Speaker, “Evaluating Alignment in LLMs,” AI Safety Conference 2023
- Panel Moderator, “The Future of AI Governance,” Berlin Tech Summit 2022
- Workshop Leader, “Practical Approaches to AI Alignment,” NeurIPS 2021

Languages: English (fluent), German (native), Polish (native), French (conversational)