GIADA PISTILLI

EDUCATION

Sorbonne Université (Paris, France) *Nov 2019 - Mar 2024*
Ph.D. in Philosophy
Laboratory: Sciences, Norms, Democracy (UMR 8011) – Sorbonne Université & CNRS
Dissertation Topic: *For an Ethics of Conversational Artificial Intelligence*
- Supervisor: Anouk Barberousse
- Thesis Committee: Seth Lazar, Raja Chatila, Mickaël Cozic, Isabelle Drouet, Denis Bonnay

Sorbonne Université (Paris, France) *Sept 2017 - Sept 2019*
Master's in Philosophy, Ethics and Political Philosophy, graduated with highest honors
Dissertation Topic: *Artificial Intelligence for the Common Good: Participatory Democracy and Emerging Technologies*
- Supervisor: Serge Audier

Sciences Po (Rennes, France) *Aug 2012 - Aug 2013*
Certificate of Political Studies – Erasmus Program in Political Science and International Relations

Università degli Studi di Teramo (Teramo, Italy) *Sept 2011 - Dec 2015*
Bachelor's in Political Science and International Relations, graduated with highest honors

WORK EXPERIENCE

Hugging Face 05/2022 - today
Principal Ethicist *Paris, France*

· Led high-impact research on AI ethics, evaluation techniques, and data governance, from conception to publication in top-tier venues, collaborating with researchers, engineers, and policy experts.
· Drafted and implemented ethical frameworks for Large Language Models (LLMs) and diffusion models into system design, collaborating cross-functionally to ensure ethical compliance.
· Led Hugging Face Hub's content policy development, creating and maintaining policy documents in collaboration with engineering, legal, and policy teams.
· Co-chaired the Legal & Ethical Scholarship team for BigScience, coordinating the drafting of the ethical charter across 1000+ researchers from 60+ countries.
· Advised internal teams and external collaborators on responsible AI deployment, focusing on safety protocols and compliance across the Hugging Face Hub.
· Consulted with EU institutions and local governments on AI policy, offering expert input on content moderation and contributing to discussions on AI regulation and governance.
· Synthesized complex sociotechnical considerations into practical ethical recommendations for diverse audiences through conferences, keynotes, and interviews.

Polaria - ex Les Petits Bots 10/2019 - 05/2022
Principal Research Engineer in Ethics *Paris, France*

· Led an interdisciplinary team of software engineers and designers in developing and implementing conversational agents for French local governments, as part of doctoral research. Successfully defended PhD thesis on this work.
· Conducted UX and UI research (e.g., focus groups, interviews, user studies) to guide the responsible design and deployment of public sector conversational agents, impacting over 1M+ users. Analyzed user interactions and feedback to continuously improve ethical standards and user experience.

· Contributed to the development of chatbots' knowledge bases, applying natural language and dialogue design techniques to optimize user interactions. Structured information to improve conversational flow and user engagement.
· Researched ethical challenges around AI anthropomorphization, data protection, and compliance, translating insights into actionable guidelines for development.
· Led monthly ethics workshops to address real-time ethical challenges in AI, promoting a culture of responsible innovation across the company.
· Developed and established ethical review processes for AI projects, including drafting the company's ethical charter for customer contracts, conforming to legal standards and internal ethical principles.
· Gained on-field, hands-on experience in AI deployment, directly overseeing the implementation and operation of conversational agents in diverse local government settings.

ESILV Léonard de Vinci – École d'Ingénieurs 09/2021 - 05/2022
Lecturer *Paris, France*

· Taught Ethics, Engineering Ethics, and Digital Ethics to international master's engineering students, communicating complex ethical concepts in technical contexts.

European Parliament 02/2016 - 08/2016
Policy Advisor Trainee *Brussels, Belgium*

· Served as policy advisor for the European Parliament's Subcommittee on Human Rights (DROI) , focusing on international human rights and EU affairs.
· Drafted, refined, and proposed amendments to bills on corporate social responsibility.
· Synthesized geopolitical insights into accessible briefing materials for Members of the European Parliament (MEPs).
· Represented the subcommittee at meetings and conferences within the European Parliament, engaging with various stakeholders on sensitive human rights issues (e.g., United Nations).
· Collaborated on developing voting recommendations strategies for plenary sessions in Strasbourg, integrating diverse perspectives on international human rights policies.

Centre d'Études Européennes, Recherches et Nouvelles Technologies 04/2014 - 10/2014
Young Researcher *Brussels, Belgium*

· Co-authored "Guida Horizon alla salute", a collaborative guide on European medical research policy, contributing to the synthesis of regulatory considerations for a medical audience.
· Contributed to drafting "CitiTrust", a European project aimed at promoting citizen engagement and fostering collaboration across diverse communities in Europe.
· Represented the association at European Commission meetings, gaining insight into EU-level policy-making processes and stakeholder engagement in research contexts.

SELECTED PUBLICATIONS

- Johnson, R.L., **Pistilli, G.**, et al. (2022). "The Ghost in the Machine has an American accent: value conflict in GPT-3". Available here.
  - Investigated cultural value alignment in GPT-3, revealing biases towards US-centric values. Explored how language and cultural values impact large language models, and proposed recommendations for addressing value pluralism in AI systems.

- **Pistilli, G.**, Munoz Ferrandis, C., Jernite, Y., Mitchell, M. (2023). "Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML". In FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 343-354. Available here.
  - Analyzed the interdependence of ethics, law, and computer science in AI governance. Proposed

collaborative governance tools to shape positive AI development and deployment, emphasizing the role of ethical charters, licenses, and model cards.

- **Pistilli, G.**, Leidinger, A., Jernite, Y., Kasirzadeh, A., Luccioni, A.S., Mitchell, M. (2024). "CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models". In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 1132-1144. Available here.
  - Developed CIVICS, a multilingual dataset to evaluate social and cultural variations in LLMs across value-sensitive topics. Investigated LLMs' responses to ethical issues through experiments, revealing variability across models and languages.

- Rocca, R., **Pistilli, G.**, Maheshwari, K., Fusaroli, R. (2024). "Introducing ELLIPS: An Ethics-Centered Approach to Research on LLM-Based Inference of Psychiatric Conditions". In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 1243-1254. Available here.
  - Developed ELLIPS, an ethical toolkit for guiding research on language-based inference of psychopathology using LLMs. Identified seven core ethical principles and translated them into practical questions for researchers, aiming to facilitate the development of clinically applicable AI models in mental health care.

- Akiki, C., **Pistilli, G.**, Mieskes, M., Gallé, M., Wolf, T., Ilic, S., Jernite, Y. (2022). "BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model." In NeurIPS 2022 Workshop on Broadening Research Collaborations. Available here.
  - Explored the interdisciplinary collaboration behind the creation of ROOTS, a large multilingual dataset for training BLOOM. Analyzed challenges and successes in leading large-scale participatory research and outlined the social impact beyond technical outputs.

SKILLS

Languages:

- Italian - Native proficiency
- French - Bilingual proficiency
- English - Full professional proficiency
- Spanish - Professional working proficiency

Technical AI Knowledge:

- Expertise in conversational AI, including design, implementation, and ethical considerations of interactive language systems.
- Knowledge of AI architectures including LLMs, transformer architectures, and generative AI approaches.
- Experience in AI development processes, including ethical implementation, data governance, and deployment (e.g., BigScience project).
- Proficiency in applying AI ethics frameworks to conversational AI and interactive systems.
- Understanding of sociotechnical challenges in AI, particularly multilingualism and evaluation techniques (e.g., ROOTS dataset, CIVICS project).