# Responsible AI in the Era of Large Language Models

Christos Christodoulopoulos[1]

[1]Amazon, UK

## 1. Abstract

Large Language models are now ubiquitous, and since the release of ChatGPT last November, are no longer an academic curiosity. As LLMs become part of products used daily by millions of people, there is an increased urgency to ensure that these models are developed and operate responsibly. In this talk, I am going to discuss the what Responsible AI (RAI) looks like in this new era, how RAI is practiced in an industry setting and how it is influenced by and inspires foundational research into RAI topics. I will talk about two recently-published projects from my team that cover two of the many topics associated with RAI. Looking at Fairness, I will present TANGO, a new dataset that measures Transgender and Nonbinary biases in open language generation. In the area of Privacy, I will present a method for controlling the memorisation of potentially sensitive training data through prompt tuning. I will conclude with a look at the use of such RAI research in practice and examples of RAI mitigation strategies for production-ready LLMs.

## 2. Short Biography

Dr Christos Christodoulopoulos is a Senior Applied Scientist at Amazon, currently working on Responsible AI for Alexa and LLMs. He was previously part of the Alexa AI Knowledge team, working on entity linking and relation extraction for Knowledge Graph-based question answering. He got his PhD at the University of Edinburgh, where he studied the underlying structure of syntactic categories across languages. Before joining Amazon, he was a postdoctoral researcher at the University of Illinois working on constraint-based inference for semantic role labeling and psycholinguistic models of language acquisition. He is an editor for the Northern European Journal Journal of Language Technology, an area chair for a number of *CL conferences, and the general chair for the 2021 Truth and Trust Online conference.

**Personal Website.** https://christos-c.com/

*NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence, November 6–7, 2023, Rome, Italy*

chrchrs@amazon.co.uk (C. Christodoulopoulos)