ŞTEFAN SARKADI

PhD Candidate in Artificial Intelligence

@ stefansarkadi@gmail.com London, UK www.stefansarkadi.com

RESEARCH EXPERIENCE

PhD Research

King's College London, Dept. of Informatics

Oct 2016 – Present London,UK

- Research, design, implementation and evaluation of Agent Based Models and Multi-Agent Systems.
- Engineering of complex reasoning agents and communication protocols using Knowledge Engineering techniques.
- Extensive interdisciplinary research on the topic of machine deception using a holistic approach covering literature from Psychology, Philosophy, Sociology, Economics, Neuroscience and Communication Theory.

Visiting PhD Research

MIT, Media Lab

Jul 2018 – Oct 2018 Cambridge, MA

- Research, design, implementation and evaluation of evolutionary game-theoretical models of agents.
- Development of evolutionary models using high-level cognitive architectures to promote cooperation and ethical behaviour in agent societies.

Research Assistant

King's College London, Dept. of Informatics

Sep 2015 – Sep 2016 London, UK

- Research on the feasability of applying Blockchain technology for Nuclear Non-Proliferation.
- Big Data analysis of wheat market data for the development of market behaviour models.

TEACHING EXPERIENCE

Associate Fellow

The Higher Education Academy UK

2019 - present UK

Graduate Teaching Assistant

King's College London, Dept. of Informatics

Sep 2016 – Present London, UK

- Gave a guest lecture for the Artificial Intelligence module to a group of more than 150 students.
- Taught small group tutorials and seminars of 10-15 undergraduate students for: Introduction to Artificial Intelligence; Elementary Logic and Applications; Philosophy & Ethics of AI.
- Taught large group tutorials and seminars of 100 - 400 undergraduate and postgraduate students for: Artificial Intelligence; Elementary Logic and Applications; Philosophy & Ethics of AI.
- Taught and supervised lab practicals of 30-50 undergraduate and postgraduate students for: Artificial Intelligence; Machine Learning; Computer Programming for Data Science.

RESEARCH INTERESTS

- **Deceptive AI.**
- Agents and multi-agent systems.
- Ethics of AI.
- XAI: Storytelling and Argumentation.

EDUCATION

MSc. in Mind, Language, and Embodied Cognition (Cognitive Science)

The University of Edinburgh

2014 - 2015 Edinburgh, Scotland

B.A. (with Hons.) in Philosophy

West University of Timisoara

2011 - 2014 Timisoara, Romania

MAIN PROJECTS

Deceptive Machines

- This is my main project; it looks into the ways in which we can model, design and engineer machines that can deceive and that can detect deception. The short-term aim of this project is to model deceptive interactions between artificial agents. The long-term aim of this project is to understand how to prevent and mitigate the malicious behaviour of machines that, in the future, might develop their own reasons to deceive.

Artificial Theory of Mind

- This project, which is closely tied to the one mentioned above, looks into the ability of artificial agents to model the minds of other agents (human or artificial). This ability enables machines to deceive by manipulating the beliefs of their targets. The aim of this project is to understand how machines might form Theories-of-Mind of their targets through communication.

Machine Behaviour & Society

- This project looks into how machines that exhibit unethical behaviour, such as deception, impact society. Some questions this project aims to answer are: What are the ethical implications of deceptive machines? Is there ethical machine deception? What forms can machine deception take in society? Are Computer Science techniques appropriate for the study of machine deception, or any other type of machine behaviour?

ACADEMIC SERVICE

1st International Workshop on Deceptive AI (DeceptECAI 2020) @ECAI2020

Co-Chair

2020 Santiago de Compostela

Online Handbook for Argumentation in Artificial Intelligence (OHAAI)

Co-Founder & Editor

International Workshop on Explainable Transparent Autonomous Agent and Multi-Agent Systems (EXTRAAMAS)

PC Member

Annual International Conference on Human-Agent Interaction (HAI)

Reviewer

Argumentation Reading Group King's College London

Co-Founding Member

Journal of Logic and Computation (JLC)

Reviewer

The Knowledge Engineering Review (KER)

Reviewer

AWARDS & GRANTS

- Online Deception Survey Research Grant, Alan Turing Institute ARC (2020).
- Nominated for KCL Dept. of Informatics *Outstanding Teaching Assistant Award* (2018,2019).
- Two *Best Early Researcher Paper* nominations at the EUMAS-AT conference (2018).
- *Graduate Visiting Researcher Funding*, MIT Media Lab (2018).
- *Conference Travel Grant for IJCAI '18*, Artificial Intelligence Journal (2018).
- *NMS Faculty Studentship Scheme*, King's College London (2018-2020).
- *Graduate Teaching Studentship*, King's College London (2016-2018).
- *Academic Performance Scholarship*, West University of Timişoara (2012-2014).

SKILLS

LaTeX
Python, R

LANGUAGE SKILLS

English
Romanian
German
French
Italian

TALKS & LECTURES

AI & Ethics

Guest Lecture for the Artificial Intelligence Module, King's College London

Dec 2019 London, UK

Deceptive Autonomous Agents

Shrivenham Defence and Security Symposium

Nov 2019 Shrivenham, UK

Superintelligence

Seminar for Philosophy & Ethics in AI Module, King's College London

Nov 2019 London, UK

Artificial Consciousness

Seminar for Philosophy & Ethics in AI Module, King's College London

Oct 2019 London, UK

Deceptive Storytelling in Argumention Games

Reasoning and Planning Group Seminar, King's College London

May 2019 London, UK

Deceptive Storytelling in Artificial Dialogue Games

AAAI 2019 Spring Symposium

March 2019 — Stanford, California

Towards an Approach for Modelling Uncertain Theory of Mind in Multi-Agent Systems

EUMAS-AT 2018

Dec 2018 — Bergen, Norway

On the Formal Semantics of Theory of Mind in Agent Communication

EUMAS-AT 2018

Dec 2018 — Bergen, Norway

Lies, Bullshit and Deception in Agent-Oriented Programming Languages

20th International TRUST Workshop @ IJCAI/AAMAS

July 2018 — Stockholm, Sweden

Is Your AI Cheating on You?

Doctoral Consortium of IJCAI'18

July 2018 — Stockholm, Sweden

Deception: A Multi-Agent Systems Approach

Guest Lecture for Adv. Topics in CompSci Module, King's College London

Nov 2017 — London, UK

Modelling Deception

Agents and Intelligent Systems PhD Symposium, King's College London

Aug 2017 — London, UK

Religion in the Public Cybersphere of Social Machines

COMSYMBOL 2016

Nov 2016 — Montpellier, France

Introduction to Cognitive Science

Guest Lecture for the Psychology Module, Dept. of Philosophy, West University of Timisoara

Jan 2016 — Timisoara, Romania

PUBLICATIONS

Journal Articles

- Sarkadi, Ştefan et al. (2019). "Modelling deception using theory of mind in multi-agent systems". In: *AI Communications* 32.4, pp. 287–302.

Conference Proceedings

- Mosca, Francesca et al. (2020). "Agent EXPRI: Licence to Explain". In: *Proc. of the 2nd International Workshop on EXplainable TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*. Auckland, New Zealand: Springer.
- Sarkadi, Stefan (2020b). "Deceptive Autonomous Agents". In: *Proc. of the Defence and Security Doctoral Symposium at Shrivenham*. Cranfield University.

- Sarkadi, Stefan, Peter McBurney, and Simon Parsons (2019). "Deceptive Storytelling in Artificial Dialogue Games". In: *Proc. of the AAAI 2019 Spring Symposium Series on Story-Enabled Intelligence*. Stanford, CA.
- Panisson, Alison R. et al. (2018a). "Lies, Bullshit, and Deception in Agent-Oriented Programming Languages". In: *Proc. of 20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018)*. CEUR-WS. Stockholm, Sweden, pp. 50–61.
- – (2018b). "On the Formal Semantics of Theory of Mind in Agent Communication". In: *Proc. of 6th International Conference on Agreement Technologies*. Bergen, Norway: Springer.
- Sarkadi, Stefan (2018). "Deception". In: *Proc. of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, pp. 5781–5782.
- Sarkadi, Stefan, Alison R. Panisson, et al. (2018). "Towards an Approach for Modelling Uncertain Theory of Mind in Multi-Agent Systems". In: *Proc. of 6th International Conference on Agreement Technologies*. Bergen, Norway: Springer.

Edited Books

- *Online Handbook of Argumentation for AI* (2020). Vol. 1. In Press. ArXiv.
- *Proceedings of the First International Workshop on Deceptive AI (DeceptECAI2020)* (2020). Upcoming. Springer.

Book Chapters

- Sarkadi, Stefan (2020a). "Argumentation-based Dialogue Games for Modelling Deception". In: *Online Handbook of Argumentation for AI*. vol. 1. ArXiv, pp. 38–42.
- Lobont, Florin and Ștefan Sarkadi (2016). "Religion in the public cybersphere of social machines". In: *ComSymbol 2016: Religion(s), Laïcité(s) Et Société(s) Au Tournant Des Humanités Numériques*. Ed. by Mihaela-Alexandra Tudor and Stefan Bratosin.
- Sarkadi, Ștefan (2016). "Artificial Consciousness in an Artificial World". In: *Communication Today: An Overview from Online Journalism to Applied Philosophy*. Ed. by M. Micle and C. Mesaroș. Trivent, p. 777.