STEFAN SARKADI

Researcher in Artificial and Human Intelligence

@ stefan.sarkadi@inria.fr London, UK www.stefansarkadi.com

SUMMARY

- I am a researcher in Artificial Intelligence whose perspective is highly interdisciplinary. This enables me to develop research approaches by integrating methodologies from Computer Science, Philosophy, Psychology, Sociology, Intelligence Analysis, and Communication Theory.
- My interests revolve around the topics of deception and deception detection, self-explainable AI agents with Theory-of-Mind, and the ability of AI agents to build stories and narratives. More generally, I wish to understand the behaviour of intelligent agents (humans or machines) inside social environments like hybrid societies, and study their behaviour with respect to social norms and to ethical, legal, and safety standards.
- Areas of research: Deception; Agent-based Modelling & Multi-Agent Systems; AI Ethics; Explainable AI.

EDUCATION

PhD in Computer Science (Artificial Intelligence)

King's College London

2016 - 2020 London, UK

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

The University of Edinburgh

2014 - 2015 Edinburgh, UK

B.A. (with Hons.) in Philosophy

West University of Timisoara

2011 - 2014 Timisoara, Romania

RESEARCH EXPERIENCE

Post-Doctoral Research Fellow

INRIA, Sophia-Antipolis

Nov 2020 – present Sophia-Antipolis, France

- 3IA Côte d'Azur Project: Design and test a multi-agent models and protocols to orchestrate the interactions between agents that embed different AI methods. The aim of this project is ensuring an optimized collaboration to augment, improve, and govern knowledge sharing activities in Multi-Agent Systems.
- Hyper-Agents Project: Define a new class of Multi-Agent Systems that use hypermedia as a general mechanism for uniform interaction to support AI interoperability and traceability in complex interconnected systems.

PhD Researcher

King's College London, Dept. of Informatics

Oct 2016 – Oct 2020 London, UK

- Thesis title: *Deception*
- Research, design, implementation and evaluation of Agent Based Models and Multi-Agent Systems for the study of deception.
- Engineering of complex reasoning and communication mechanisms for deceptive artificial agents using techniques from Game-Theory, Knowledge Representation and Agent-Oriented Programming Languages.
- Extensive interdisciplinary research on the topic of human and machine deception covering literature from Psychology, Philosophy, Sociology, Economics, Neuroscience and Communication Theory.

Visiting PhD Researcher

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 where deception is present.

Research Assistant

King's College London, Dept. of Informatics

Sep 2015 – Sep 2016 | London, UK

- Research on the feasability of applying Blockchain technology for non-proliferation and arms control.
- Big Data analysis of wheat market data for the development of market behaviour models.

TEACHING EXPERIENCE

Associate Fellow of the HEA

The Higher Education Academy UK

2019 - present | London, UK

Graduate Teaching Assistant

King's College London, Dept. of Informatics

Sep 2016 – Dec 2019 | London, UK

- Gave a guest lecture on *Ethics and AI* for the Artificial Intelligence module to a group of more than 150 students.
- Gave a guest lecture on how to represent deception in multi-agent systems for the Advanced Research Topics in Computer Science module to a group of more than 50 postgraduate students.
- Marked undergraduate and postgraduate courseworks for the Artificial Intelligence module.
- Taught small group tutorials and seminars of 10-15 undergraduate students for: *Introduction to Artificial Intelligence*; *Elementary Logic and Applications*.
- Taught large group tutorials and seminars of 50 - 300 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*; *Introductory Course to Python* for the MSc in Data Science.

AWARDS & GRANTS

- Online Deception Survey Research Grant, The Alan Turing Institute defence and security ARC (2020). I was lead researcher. Grant total: **£ 8960**.
- Nominated as a Graduate Teaching Assistant for the university-wide *2020 King's Education Awards*, at King's College London.
- Nominated for the Department of Informatics *2018/2019 Outstanding Teaching Assistant Award*, at King's College London.
- 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).

PROFESSIONAL ACTIVITIES

- *Co-chair* of the 1st and 2nd International Workshops on Deceptive AI @ECAI2020 & @IJCAI2021.
- *Co-founder and co-editor* of the Online Handbook of Argumentation for Artificial Intelligence (OHAAI).
- Responsible for organising and co-hosting the Distributed AI research group seminars in the Dept. of Informatics at King's College London.
- *Co-founder* of the Argumentation Reading Group at King's College London.
- *PC Member* - International Workshop on Explainable Transparent Autonomous Agent and Multi-Agent Systems (EXTRAAMAS).
- *Reviewer* - Annual International Conference on Human-Agent Interaction (HAI 2018).
- *Reviewer* - Journal of Autonomous Agents & Multi-Agent Systems (JAAMAS).
- *Reviewer* - Decision Support Systems (DECSUP).
- *Reviewer* - Journal of Logic and Computation (JLC).
- *Reviewer* - The Knowledge Engineering Review (KER).

PUBLICATIONS

Journals

Ștefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, Martin Chapman [2019]: *Modelling Deception using Theory of Mind in Multi-Agent Systems*. In: *AI Communications* 32.4, pp.287–302.

Conference Proceedings

Mosca, Francesca, **Ștefan Sarkadi**, Jose M. Such, Peter McBurney [2020]: Agent EXPRI: Licence to Explain. *Proceedings of 2nd International Workshop on EXplainable TRansparent Autonomous Agents and Multi-Agent Systems*, Auckland, New Zealand, 9-13 May 2020.

Ștefan Sarkadi [2019]: Deceptive Autonomous Agents. *Proceedings of the Shrivenham Defence and Security Doctoral Symposium*, Shrivenham, UK, 12-13 Nov 2019.

Ștefan Sarkadi, Peter McBurney, Simon Parsons [2019]: Deceptive Storytelling in Artificial Dialogue Games. *Proceedings of the AAAI 2019 Spring Symposium on Story-Enabled Intelligence*, Stanford, USA, 25-27 March 2019.

Ștefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons [2018]: Towards an Approach for Modelling Uncertain Theory of Mind in Multi-Agent Systems. *Proceedings of the 6th International Conference on Agreement Technologies*, Bergen,Norway, 6-7 December 2018.

Alison R. Panisson, **Ștefan Sarkadi**, Peter McBurney, Simon Parsons, Rafael H. Bordini [2018]: On the Formal Semantics of Theory of Mind in Agent Communication. *Proceedings of the 6th International Conference on Agreement Technologies*, Bergen,Norway, 6-7 December 2018.

Ștefan Sarkadi [2018]: Deception. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI 2018, Stockholm, Sweden, 13-19 July 2018.

Alison R. Panisson, **Ștefan Sarkadi**, Peter McBurney, Simon Parsons, Rafael H. Bordini [2018]: Lies, B*llshit, and Deception in Agent-Oriented Programming Languages. *Proceedings of the 20th International TRUST Workshop (TRUST 2018)*, IJCAI 2018, Stockholm, Sweden, 14/15 July 2018.

Edited Collections

Online Handbook of Argumentation for AI [Upcoming 2021]. Vol.2. ArXiv.

Proceedings of the First International Workshop on DeceptiveAI (DeceptECAI 2020). Upcoming. Springer.

Online Handbook of Argumentation for AI [2020]. Vol.1. ArXiv.

Book Chapters

Ștefan Sarkadi [2020]: Argumentation-based Dialogue Games for Modelling Deception. In: Online Handbook for Argumentation in AI Vol.1.

Florin Lobonț, **Ștefan Sarkadi** [2016]: Religion in the public cybersphere of social machines. 3e Colloque International Comsymbol (Comsymbol 2016), Montpellier, France, 9-10 Nov 2016. Book Chapter in Mihaela-Alexandra Tudor and Stefan Bratosin (Eds.): Religion(s), Laïcité(s) Et Société(s) Au Tournant Des Humanités Numériques.

Ștefan Sarkadi [2016]: Artificial Consciousness in an Artificial World. In: M. Micle and C. Mesaroș (Edis): *Communication Today: An Overview from Online Journalism to Applied Philosophy*, Trivent Publishing.

TALKS & LECTURES

Deceptive AI

WIMMICS Seminar Series, Inria

Feb 2021 — Sophia-Antipolis, France

Deceptive AI

Distributed AI Seminar Series, King's College London

Nov 2020 — London, UK

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

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 Advanced Research Topics in Computer Science 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

Blockchains for Non-Proliferation and Arms Control

Poster Talk at Big Data Day @King's, King's College London

Oct 2016 — London, UK

Introduction to Cognitive Science

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

Jan 2016 — Timisoara, Romania