ŞTEFAN SARKADI

Researcher in Artificial Intelligence

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

RESEARCH INTERESTS

My research interest are rooted in the areas of Multi-Agent Systems, AI & ethics, and Explainable AI. I am particularly interested in machine deception, in self-explainable agents that have the ability to model other minds, and in the relation between argumentation and storytelling. More generally, I am interested how these aspects of AI might influence the behaviour of machines as actors in hybrid societies.

Deceptive Machines

- Looking into the ways in which we can model, design and engineer machines that can deceive and that can detect deception. The modelling of deceptive interactions between artificial agents can inform us about the mechanics of deception. In the long-term, we might be able to sufficiently understand deceptive communication in order to prevent and mitigate the malicious behaviour of artificial agents that, in the future, might even develop their own reasons to deceive.

Explainable AI and Theory of Mind

- Exploring the ability of artificial agents to model the minds of other agents (human or artificial). The aim is to understand how machines might form Theories of Mind of their interlocutors through communication. Theory of Mind has the potential to increase the social ability of artificial agents, by enabling them to give better explanations that they generate by taking into account the beliefs of their interlocutors.

Machine Behaviour & Society

- Studying how machine behaviour impacts society from a anthropological and socio-economical perspectives using methods such as agent-based modelling. Some pertinent questions coming from this perspective are: What are the ethical implications of machine behaviour? How can we ensure machines behave ethically and not maliciously? What other methodologies can we develop to successfully study the behaviour of artificial agents inside complex social systems, such as hybrid societies?

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

SKILLS

- **Programming Languages**: Python, R, LaTeX, HTML, CSS.
- **Natural Languages**: English (fluent), Romanian (fluent), German (upper intermediate), French (intermediate), Italian (basic-intermediate).

RESEARCH EXPERIENCE

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.
- Engineering of complex reasoning agents and communication protocols using Knowledge Engineering techniques and Agent-Oriented Programming Languages.
- 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 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.
- 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 - 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.

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

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.

AWARDS & GRANTS

- Online Deception Survey Research Grant, The Alan Turing Institute defence and security ARC (2020). I was lead researcher. Grant total: **£ 8960**.
- 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).

ACADEMIC SERVICE

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

Co-Chair

2020 Santiago de Compostela

Online Handbook of Argumentation for 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

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

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

REFEREES

Prof. Peter McBurney
Professor of Artificial Intelligence, Dept. of Informatics, King's College London, UK
e-mail: peter.mcburney@kcl.ac.uk

Prof. Simon Parsons
Global Professor of Machine Learning, School of Computer Science, University of Lincoln, UK
email: sparsons@lincoln.ac.uk

Dr. Rafael H. Bordini
Associate Professor, FACIN-PUCRS, Brazil
email: rafael.bordini@pucrs.br

Dr. Alex Rutherford
Senior Research Scientist, Max-Planck Institute for Human Development, Germany
email: rutherford@mpib-berlin.mpg.de