

# DR. ȘTEFAN SARKADI

Researcher in Artificial Intelligence

## SUMMARY

---

- I am a researcher in Artificial Intelligence with a highly interdisciplinary perspective. This enables me to develop approaches by integrating methodologies from Computer Science, Philosophy, Psychology, Sociology, Intelligence Analysis, and Communication Theory.
- Generally, I investigate the reasoning and behaviour of intelligent agents in hybrid societies. In particular, I am interested in aspects of knowledge-sharing interactions such as reasoning, communication, information dissemination, and explanation, in heterogenous Self-organising and Open Multi-agent Systems. I am also curious to understand the legal, ethical, and safety aspects of these social interactions, with a particular focus on deception.
- Areas of research: Deception; Multi-Agent Systems; Knowledge Representation; AI Ethics; Explainable AI.

## EDUCATION

---

PhD in Computer Science (Artificial Intelligence)

**King's College London**

📅 2021

📍 London, UK

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

**The University of Edinburgh**

📅 2015

📍 Edinburgh, UK

B.A. (with Hons.) in Philosophy

**West University of Timisoara**

📅 2014

📍 Timisoara, Romania

## RESEARCH EXPERIENCE

---

Associate Researcher

**INRIA**

📅 Jan 2022 – present

📍 London & Sophia-Antipolis, UK & France

- Working on developing open Multi-Agent Systems that allow successful knowledge-sharing interactions between different types of AI agents (Logic-based, ML, neurosymbolic). I am also part of INRIA London and the WIMMICS and Hyperagents research groups.

---

Post-Doctoral Research Associate

**King's College London, Dept. of Informatics**

📅 Jan 2022 – present

📍 London, UK

- Working on developing neuro-symbolic and socially-aware AI agents that are able to learn and reason about privacy norms - in the PRAISE project within the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN).
  - Assisting Prof. Jose Such with the supervision of a PhD student working on AI for privacy.
-

## 3IA Post-Doctoral Research Fellowship

### INRIA, Sophia-Antipolis

📅 Nov 2020 – Dec 2021

📍 Sophia-Antipolis, France

- 3IA Côte d'Azur Project: Design and test multi-agent models and protocols to orchestrate the interactions between agents that embed different AI methods. The aim of this project is ensuring an optimised collaboration to augment, improve, and govern knowledge sharing activities in Open 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 on the Web.

---

## 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 feasibility 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 QUALIFICATIONS & 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

- Guest lecturer - *Artificial Intelligence*.
- Guest lecturer - *Advanced Research Topics in Computer Science*.
- Design and organisation of tutorials and seminar materials - *Philosophy & Ethics of AI and Artificial Intelligence*.
- Coursework marking - *Artificial Intelligence*.
- Small group tutorials and seminars teaching - *Introduction to Artificial Intelligence and Elementary Logic & Applications*.

- Large group tutorials and seminars teaching - *Artificial Intelligence; Elementary Logic & Applications; and Philosophy & Ethics of AI*.
- Lab and practicals teaching - *Artificial Intelligence; Machine Learning; Computer Programming for Data Science; Introductory Course to Python* for the MSc in Data Science.

## AWARDS & GRANTS

---

- Lead researcher for *Online Deception Survey Research Grant*, The Alan Turing Institute defence and security ARC (2020). Grant total: **£ 8960**.
- Nomination - Graduate Teaching Assistant for the university-wide *2020 King's Education Awards*, at King's College London.
- Nomination - 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 Timișoara (2012-2014).

## PROFESSIONAL ACTIVITIES

---

- *General chair* of the 1st and 2nd International Workshops on Deceptive AI @ECAI2020 & @IJCAI2021.
- *Co-chair* of the 1st International Workshop on Argumentation & Machine Learning @COMMA2022.
- *Co-founder and co-editor* of the Online Handbook of Argumentation for Artificial Intelligence (OHAAl).
- *Co-organiser* of the Distributed AI research group seminars, Dept. of Informatics, King's College London.
- *PC Member* - 1st, 2nd and 3rd International Workshops on Explainable Transparent Autonomous Agent and Multi-Agent Systems (EXTRAAMAS).
- *PC Member* - 1st International Conference on AI for People (CAIP'21).
- *PC Member* - 1st Workshop on Argumentation for eXplainable Artificial Intelligence (ArgXAI 2022).
- *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 Applied Logic (JAL).
- *Reviewer* - Journal of Logic and Computation (JLC).
- *Reviewer* - The Knowledge Engineering Review (KER).
- *Reviewer* - Philosophical Psychology.

## PUBLICATIONS

---

### Journals

Stefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, Iyad Rahwan [2021]: *The evolution of deception*. In: *Royal Society Open Science* 8:201032, <https://doi.org/10.1098/rsos.201032>.

Stefan 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., DOI: 10.3233/AIC-190615, <https://content.iospress.com/articles/ai-communications/aic190615>

### Conference Proceedings

Xiao Zhan, **Stefan Sarkadi**, Natalia Criado Pacheco, Jose Such [2022]: A Model for Governing Information Sharing in Smart Assistants. *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society*, Oxford, UK, 1-3 August 2022. (Accepted/In Press)

Ştefan 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. [https://link.springer.com/chapter/10.1007/978-3-030-17294-7\\_1](https://link.springer.com/chapter/10.1007/978-3-030-17294-7_1) **Best Early Researcher Paper Nomination**

Alison R. Panisson, Ştefan 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. [https://link.springer.com/chapter/10.1007/978-3-030-17294-7\\_2](https://link.springer.com/chapter/10.1007/978-3-030-17294-7_2) **Best Early Researcher Paper Nomination**

Ştefan Sarkadi [2018]: Deception. *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018*, Stockholm, Sweden, 13-19 July 2018. <https://www.ijcai.org/proceedings/2018/834>

---

### Workshops with Proceedings

Mosca, Francesca, Ştefan 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. [https://link.springer.com/chapter/10.1007/978-3-030-51924-7\\_2](https://link.springer.com/chapter/10.1007/978-3-030-51924-7_2)

---

### Workshops & Symposia

Ştefan Sarkadi [2019]: Deceptive Autonomous Agents. *Proceedings of the Shrivenham Defence and Security Doctoral Symposium*, Shrivenham, UK, 12-13 Nov 2019. [https://cord.cranfield.ac.uk/articles/presentation/Deceptive\\_Autonomous\\_Agents/11558397](https://cord.cranfield.ac.uk/articles/presentation/Deceptive_Autonomous_Agents/11558397)

Ştefan 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. <http://logical.ai/story/papers/sarkadi.pdf>

Alison R. Panisson, Ştefan 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. <http://ceur-ws.org/Vol-2154/paper5.pdf>

---

### Edited Collections

Ştefan Sarkadi, Benjamin Wright, Peta Masters, Peter McBurney [2021]. *Deceptive AI*. Springer CCIS vol.1296. <https://doi.org/10.1007/978-3-030-91779-1>

*Online Handbook of Argumentation for AI* [2021]. Vol.2. ArXiv. <https://arxiv.org/abs/2106.10832>

*Online Handbook of Argumentation for AI* [2020]. Vol.1. ArXiv. <https://arxiv.org/abs/2006.12020>

---

### Book Chapters

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

Florin Lobont, Ştefan 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*.

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

# TALKS & LECTURES

---

How to Model Deception with AI

**Deception and Autonomous Systems Seminar Series, WASP-HS**

📅 May 2022

📍 Sweden, (online)

Trustworthy deceptive machines

**Guest Lecture for Trust Systems Course, Faculty of Business & IT, Ontario Tech University**

📅 Mar 2022

📍 Ontario, Canada (online)

Creating deceptive machines

**UK National Cyber Deception Symposium**

📅 Dec 2021

📍 Shrivenham, UK (online)

A Short Introduction into Deception & AI

**Philosophy of AI Seminar Series, Dept. of Philosophy, West University of Timisoara**

📅 May 2021

📍 Timisoara, Romania (online)

AI & Deception in Multi-Agent Systems

**WIMMICS Seminar Series, Inria**

📅 Feb 2021

📍 Sophia-Antipolis, France (online)

Deceptive AI

**Distributed AI Seminar Series, King's College London**

📅 Nov 2020

📍 London, UK (online)

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 Argumentation 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

## Introduction to Cognitive Science

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

📅 Jan 2016

📍 Timisoara, Romania