

# DR. ȘTEFAN SARKADI

Proleptic Lecturer (Assistant Professor)

RAEng UK Intelligence Community Research Fellow

 <https://www.stefansarkadi.com/>

## SUMMARY

---

- As an interdisciplinary researcher in Artificial Intelligence I aim to deliver highly impactful research at the intersection of Computer Science, Cognitive Science, Philosophy, and Intelligence Analysis.
- Generally, I investigate the reasoning and behaviour of intelligent agents in hybrid societies, which are societies where human and AI agents interact. In particular, I am interested in aspects of knowledge-sharing interactions such as reasoning, communication, information dissemination, and explanation, in complex adaptive and open multi-agent systems. My main focus and expertise is in the modelling, analysis, and explanation of deception and deceptive AI.
- Areas of research: Deception; Multi-Agent Systems; Intelligence Analysis; 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

---

Proleptic Lecturer (Assistant Professor) & RAEng UK IC Research Fellow

King's College London

 Nov 2023 – present

 London, UK

- Responsible for conducting and leading impactful research in AI and writing research grant proposals.
- Supervision of PhD and MSc students on AI-related topics.
- My research tackles the integration of techniques from AI and deception analysis to understand complex self-organising and self-adaptive human-AI societies. My collaborators include international academic research groups as well as the Intelligence & Defense communities.

---

Research Fellow

King's College London

 Dec 2022 – present

 London, UK

- Funded by the Royal Academy of Engineering UK Intelligence Community Fellowship
- £200.000 individual grant awarded for at least 2 years to explore the topic *Enhancing deception analysis with storytelling AI*
- Working on integrating techniques from AI and deception analysis to generate narratives about multi-agent interactions in complex systems in order to help intelligence analysts perform inference to the best explanation.

- My fellowship also involves international research collaborations with a potential for very high impact in International Security & Defense, which I will conduct at various institutions and groups in addition to KCL, including the National Police AI Lab and the Uni. of Utrecht in the Netherlands, Inria, CNRS, and 3iA Côte d'Azur in France, and with the Intelligence Community.
- 

## Associate Researcher

### Inria

📅 Jan 2022 – present

📍 Sophia-Antipolis, France

- Working on open multi-agent systems that allow successful knowledge-sharing interactions between different types of AI agents (Logic-based, ML, neurosymbolic). I am part of the Wimmics and Hyperagents research groups.
- 

## Post-Doctoral Research Associate

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

📅 Jan 2022 – Jan 2023

📍 London, UK

- Worked on 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, and Communication Theory.
- 

## Visiting PhD Researcher

### MIT, Media Lab

📅 Jul 2018 – Oct 2018

📍 Cambridge, MA, USA

- Research, design, implementation and evaluation of evolutionary game-theoretical models of agents.
- Development of evolutionary models using high-level cognitive architectures to study the evolution of deception.

# TEACHING QUALIFICATIONS & EXPERIENCE

---

Associate Fellow of the HEA

The Higher Education Academy UK

📅 2019 - present

📍 London, UK

---

Lecturer

King's College London, Dept. of Informatics

📅 Sep 2022 – present

📍 London, UK

- Delivered 20% of the lectures for the *Agents & Multi-Agent Systems Module* (6CCS3AMS/7CCSMAMS).
- Responsible for designing parts of the core material and designing and marking parts of the exam questions for 6CCS3AMS and 7CCSMAMS.
- Supervised 6 MSc students on topics related to the modelling of artificial societies and social simulation.

---

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

---

- RAEng UK IC Research Fellowship (2022-2024) Grant total: **£ 200,000**.
- Inria Recipient of the 3IA Postdoctoral Fellowship (2020-2022).
- Lead researcher for *Online Deception Survey Research Grant*, The Alan Turing Institute defence and security ARC (2020). Grant total: **£ 8,960**.
- 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.
- *Best Visionary Paper* selection at AAMAS (2023).
- 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 (OHAAI).
- *Session Chair* - 16th International Conference on Agents and Artificial Intelligence (ICAART 24).
- *PC Member* - 26th and 27th European Conference on Artificial Intelligence (ECAI 23 and 24).
- *PC Member* - 32nd International Joint Conference on Artificial Intelligence (IJCAI'23).
- *PC Member* - 1st, 2nd and 3rd International Workshops on Explainable Transparent Autonomous Agent and Multi-Agent Systems (EXTRAAMAS).
- *PC Member* - 1st & 2nd International Conference on AI for People (CAIP'21).
- *PC Member* - 1st Workshop on Argumentation for eXplainable Artificial Intelligence (ArgXAI 2022).
- *PC Member* - International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE 2024).
- *Reviewer* - Autonomous Agents and Multi-Agent Systems (AAMAS 2024).
- *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.
- *Reviewer* - ACM Transactions on Autonomous & Adaptive Systems (ACM TAAS).
- *Reviewer* - IEEE Transactions on Cognitive and Developmental Systems (IEEE TCDS).
- *Reviewer* - IEEE Technology & Society Magazine (IEEE TSM).

## PUBLICATIONS

---

### PhD Thesis

Stefan Sarkadi [2021]: Deception. *PhD Thesis, King's College London*, London, UK, <https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.835604>

### Journals

Stefan Sarkadi [2024]: *Self-Governing Hybrid Societies and Deception*. In: *ACM Transactions on Autonomous and Adaptive Systems*, In Press <https://doi.org/10.1145/3638549>.

Peter R. Lewis & Stefan Sarkadi [2024]: *Reflective Artificial Intelligence*. In: *Minds and Machines*, In Press [https://kclpure.kcl.ac.uk/ws/portalfiles/portal/245195444/M\\_M\\_Reflective\\_AI\\_6\\_.pdf](https://kclpure.kcl.ac.uk/ws/portalfiles/portal/245195444/M_M_Reflective_AI_6_.pdf).

Stefan Sarkadi [2023]: *Deceptive AI & Society*. In: *IEEE Technology & Society*, 42:4, pp.77-86. <https://ieeexplore.ieee.org/abstract/document/10410131>.

Stefan Sarkadi, Andreea G.B. Tettamanzi, Fabien Gandon [2022]: *Interoperable AI: Evolutionary Race Towards Sustainable Knowledge Sharing*. In: *IEEE Internet Computing* 26:6, pp. 25-32, <https://doi.org/10.1109/MIC.2022.3214378>.

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

Stefan Sarkadi & Peter R. Lewis [2024]: The Triangles of Dishonesty: Modelling the Evolution of Lies, Bullshit, and Deception in Agent Societies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Auckland, NZ, 6-10 May 2024. In Press [https://kclpure.kcl.ac.uk/ws/portalfiles/portal/245179075/CR\\_AAMAS\\_2024\\_Dishonesty\\_Triangles.pdf](https://kclpure.kcl.ac.uk/ws/portalfiles/portal/245179075/CR_AAMAS_2024_Dishonesty_Triangles.pdf)

Luis Henrique Herbets de Sousa, Guilherme Trajano, Analúcia Schiaffino Morales, **Ştefan Sarkadi**, Alison R. Panisson [2024]: Using Chatbot Technologies to Support Argumentation. In: *Proc. of 16th International Conference on Agents and Artificial Intelligence (ICAART 24)*, Feb 2024, Rome, Italy.

Milena Seibert Fernandes, Roberto Rodrigues Filho, Iwens Sene-Junior, **Ştefan Sarkadi**, Alison R. Panisson, Analúcia Schiaffino Morales [2024]: An Interpretable Machine Learning Approach for Identifying Occupational Stress in Healthcare Professionals. In: *Proc. of 16th International Conference on Agents and Artificial Intelligence (ICAART 24)*, Feb 2024, Rome, Italy.

Heitor Henrique da Silva, Michele Rocha, Guilherme Trajano, Analúcia Schiaffino Morales, **Ştefan Sarkadi**, Alison R. Panisson [2024]: Distributed Theory of Mind in Multi-Agent Systems. In: *Proc. of 16th International Conference on Agents and Artificial Intelligence (ICAART 24)*, Feb 2024, Rome, Italy.

Ştefan Sarkadi [2023]: An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies. In *Proceedings of the 4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, Toronto, Canada, 24-29 Sep 2023. In Press <https://kclpure.kcl.ac.uk/portal/en/publications/an-arms-race-in-theory-of-mind-deception-drives-the-emergence-of->

Xiao Zhan, **Ştefan Sarkadi**, Jose Such [2023]: Privacy-enhanced Personal Assistants based on Dialogues and Case Similarity. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI)*, Kraków, Poland, 2-5 Oct 2023. In Press <https://kclpure.kcl.ac.uk/ws/portalfiles/portal/224456858/264Zhan.pdf>

Ştefan Sarkadi, Peidong Mei, Edmond Awad [2023]: Should My Agent Lie for Me? A Study on Attitudes of US-based Participants Towards Deceptive AI in Selected Future-of-work Scenarios. *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, London, UK, 29 May -2 June 2023. <https://dl.acm.org/doi/abs/10.5555/3545946.3598656> **Voted among Best Visionary Papers @AAMAS**

Michele Rocha, Heitor Henrique da Silva, Analúcia Schiaffino Morales, **Ştefan Sarkadi**, Alison R. Panisson [2023]: Applying Theory of Mind to Multi-Agent Systems: A Systematic Review. *Proceedings of the Brazilian Conference in Intelligent Systems (BRACIS)*, 2023. [https://link.springer.com/chapter/10.1007/978-3-031-45368-7\\_24](https://link.springer.com/chapter/10.1007/978-3-031-45368-7_24).

Xiao Zhan, Yifan Xu, **Ştefan Sarkadi** [2023]: Deceptive AI Ecosystems: The Case of ChatGPT. *Proceedings of the 5th ACM Conference on Conversational User Interfaces (CUI)*, Eindhoven, NL, July 2023. <https://doi.org/10.1145/3571884.3603754>

Xiao Zhan, **Ştefan 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. <https://doi.org/10.1145/3514094.3534129>

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

Ştefan Sarkadi, Ionuț Moraru, Louise Manning [2023]: Sustainable AI & Agricultural Technologies. *Proc. of 1st International Workshop on Sustainable and Scalable Self-Organisation*, Toronto, Canada, Sep 2023. <https://ieeexplore.ieee.org/abstract/document/10336203>

Ştefan Sarkadi, Fabien Gandon [2023]: Interoperable AI for Self-Organisation. *Proc. of 1st International Workshop on Sustainable and Scalable Self-Organisation*, Toronto, Canada, Sep 2023. <https://ieeexplore.ieee.org/abstract/document/10336195>

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\*Ilshit, 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* [2023]. Vol.4. ArXiv. *In Press*

*Online Handbook of Argumentation for AI* [2022]. Vol.3. ArXiv. <https://arxiv.org/abs/2212.07996>

*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

Understanding Deception in Hybrid Societies

**CSIRO Data61 Human-Centric AI Seminar**

 Oct 2023

 Australia (online)

An Arms Race in Theory-of-Mind

**ACSOS 2023**

 Sep 2023

 Toronto, Canada

Working Alongside Deceptive Machines

**ICSPP DSEI 2023**

 Sep 2023

 Excel London, UK

## Deceptive AI Ecosystems

### **CUI 2023**

📅 July 2023

📍 Eindhoven, NL

### Should My Agent Lie for Me?

### **AAMAS 2023 & RAD-AI @AAMAS 2023**

📅 May 2023

📍 London, UK

### Creating Deceptive Machines

### **Monash University Cyber Security Seminar**

📅 Dec 2022

📍 Melbourne, AUS

### 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 Autonomous Agents

### **Shrivenham Defence and Security Symposium**

📅 Nov 2019

📍 Shrivenham, 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