Schedule

SIOE AI Bootcamp Schedule		
Session	**Time**	**Speaker**
Registration & Coffee	8:40 am	
Opening remarks (0.25h)	9:00 am	Gillian Hadfield
Block 1a: Tutorial Material (1.25h)	9:15 am	Introduction to ML: *Gillian Hadfield* (0.5h)
	9:45 am	Overview of AI landscape; intro to AI alignment, cooperative AI: *Lewis Hammond* (0.75h)
Break (0.25h)	10:30 am	
Block 1b: Tutorial Material (1.5h)	10:45 am	Studying the behavior of LLM-based agents in multi-agent systems: *Joel Leibo* (1h)
	11:45 am	Social choice for AI alignment: *Rachel Freedman* (0.5h)
Lunch (1h)	12:15 pm	
Block 2: Special topics (2h)	1:15 pm	Generative social dilemmas: *Dylan Hadfield-Menell*
	1:45 pm	AI and cultural evolution: *Ed Hughes*
	2:15 pm	Delegation to AI agents: *Elias Fernandez-Domingos*
	2:45 pm	Democratic AI: *Manon Revel*
Break (0.25h)	3:15 pm	
Closing remarks (0.50h)	3:30 pm	Gillian Hadfield
Program End	4:00pm	

Bios:

Elias Fernandez-Domingos is currently a Postdoctoral researcher (FNRS fellow) at the ULB – Brussels. He is interested in the origins of cooperation in social interactions and how we can maintain it in an increasingly complex and hybrid human-AI world. In his research he applies concepts and methods from (Evolutionary) Game Theory, Behavioral economics and Machine Learning to model collective (strategic) behavior and validate it through behavioral economic experiments.

Rachel Freedman is a ML PhD student at UC Berkeley, researching safe AI with the Center for Human Compatible Artificial Intelligence. She has experience in reinforcement learning, value alignment, software engineering, and is additionally interested in computational neuroscience and cognitive science. Her diverse background includes co-founding the Oxford Existential Risk Society, working in fintech in London, performing in community theatre in Mississippi, and shoveling manure in North Carolina.

Gillian Hadfield is the inaugural Schwartz Reisman Chair in Technology and Society, Professor of Law, Professor of Strategic Management at the University of Toronto and holds a CIFAR AI Chair at the Vector Institute for Artificial Intelligence. She is a Schmidt Sciences AI2050 Senior Fellow. Her research is focused on innovative design for legal and regulatory systems for AI and other complex global technologies; computational models of human normative systems; and working with machine learning researchers to build ML systems that understand and respond to human norms.

Dylan Hadfield-Menell is the Bonnie and Marty (1964) Tenenbaum Career Development Assistant Professor of EECS at MIT. He runs the Algorithmic Alignment Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and is also a Schmidt Sciences AI2050 Early Career Fellow. His research group works to address alignment challenges in multi-agent systems, human-AI teams, and societal oversight of machine learning. Their goal is to enable the safe, beneficial, and trustworthy deployment of AI in real-world settings.

Lewis Hammond is a DPhil candidate in computer science at the University of Oxford and co-director of the Cooperative AI Foundation. He is also affiliated with the Future of Humanity Institute at Oxford and the Centre for the Governance of AI. His research concerns safety and cooperation in multi-agent systems, motivated by the problem of ensuring that AI and other powerful technologies are developed and governed safely and democratically.

Ed Hughes is a Staff Research Engineer at Google DeepMind, a Visiting Fellow at the London School of Economics, and an Advisor to the Cooperative AI Foundation. Ed is a scientific leader in the field of AI and an expert on fast adaptation. His teams have pioneered large-scale reinforcement learning, the paradigm of Cooperative AI, and ad-hoc collaboration between machines and humans. He draws inspiration from diverse

sources, including cultural evolution, social psychology, economics, organizational design, and meta-learning.

Joel Leibo is a senior staff research scientist at Google DeepMind and visiting professor at King's College London. He is interested in reverse engineering human biological and cultural evolution to inform the development of artificial intelligence that is simultaneously human-like and human-compatible. In particular, Joel believes that theories of cooperation from fields like cultural evolution and institutional economics can be fruitfully applied to inform the development of ethical and effective artificial intelligence technology.

Manon Revel is a social choice theorist who researches governance in the context of human and AI decision-making. She is an Employee Fellow at the Berkman Klein Center. Her research focuses on reviving democratic governance online and offline, and understanding information disorders. Leveraging mathematical models, statistical tools and political philosophy theories, she investigates new ways to organize collectively and share information in order to update how we think and do democracy.