JUNE 1-4, 2025

CHAPMAN UNIVERSITY

Agency, Intentions, and Artificial Intelligence

INTENTIONS AND AI WORKSHOP

SPONSORED BY: Google aps Association for Psychological Science Patrick J McGovern Foundation

HOTEL:

Hyatt Regency Orange County

11999 Harbor Blvd, Garden Grove, CA 92840

(714) 750-1234

WORKSHOP LOCATION:

Killefer Conference Room A (RM 103)

Daniele C. Struppa Research Park

Chapman University

540 N. Lemon St, Orange, CA 92867

WORKSHOP CONTACTS:

Uri Maoz – maoz@chapman.edu

Tian Lan – tlan@chapman.edu

WEBSITE:

https://ai-intentions.org/events/june-2025-workshop/

Agency, Intentions, and Artificial Intelligence

AI
AI

WORKSHOP AGENDA

June 1st, 2025

Time	Event	Location
4pm	Hotel check-in available	Hyatt (Front Desk)
4pm	Workshop check-in available	Hyatt (Harbor Room)
5:30pm~9:30pm	Bar	
7pm~9pm	Group Dinner	

Agency, Intentions, and
Artificial Intelligence

AI
AI

June 2nd, 2025

Time	Session	Event	Location
7:30am		Shuttle Pick-up (to Chapman)	Hyatt (Front Entrance)
8am~9am		Breakfast	Killefer
9am~9:30am	Morning Session	Introduction	
9:30am~10:45am		Session 1	
10:45am~11:15am		Coffee Break	
11:15am~12:30pm		Session 2	
12:30pm~2:30pm		Lunch + Free Time	
2:30pm~3:45pm	Afternoon Session	Session 3	
3:45pm~4:15pm		Coffee Break	
4:15pm~5:30pm		Session 4	
5:30pm~5:45pm		Coffee Break	
5:45pm~6:15pm		Summary	
6:30pm~8:15pm		Group Dinner	Faculty Athenaeum (Argyros Forum)
8:30pm		Shuttle Pick-up (to hotel)	400 N Center St, Orange, CA 92866

Agency, Intentions, and
Artificial Intelligence

AI
AI

June 3rd, 2025

Time	Session	Event	Location
7:30am		Shuttle Pick-up (to Chapman)	Hyatt (Front Entrance)
8am~9am		Breakfast	Killefer
9am~9:30am	Morning Session	Look-Ahead	
9:30am~10:45am		Session 5	
10:45am~11:15am		Coffee Break	
11:15am~12:30pm		Session 6	
12:30pm~2:30pm		Lunch + Free Time	
2:30pm~3:45pm	Afternoon Session	Session 7	
3:45pm~4:15pm		Coffee Break	
4:15pm~5:30pm		Session 8	
5:30pm~5:45pm		Coffee Break	
5:45pm~6:15pm		Summary	
6:30pm~8:15pm		Group Dinner	O SEA (109 S. Glassell St, Orange, CA 92866)
8:30pm		Shuttle Pick-up (to hotel)	

Agency, Intentions, and
Artificial Intelligence

AI
AI

June 4th, 2025

Time	Session	Event	Location
7:30am		Shuttle Pick-up (to Chapman)	Hyatt (Front Entrance)
8am~9am		Breakfast	Killefer
9am~9:30am	Morning Session	Look-Ahead	
9:30am~10:45am		Session 9	
10:45am~11:15am		Coffee Break	
11:15am~12:30pm		Session 10	
12:30pm		Lunch + Adjourn	

Agency, Intentions, and
Artificial Intelligence

AI
AI

SESSION SCHEDULES & ROLES

June 2nd: Foundations & Empirical Grounding

Morning Sessions

Introduction (9am~9:30am)

Uri Maoz

Session 1: Defining Intention Across Domains (9:30am~10:45am)

Guiding Question: What core features define 'intention' in humans? How do those features generalize to animals and potentially to AI? How do we differentiate intentional action from reflexes or from programmed behavior (including optimization among multiple goals), and what are the philosophical and practical implications of these distinctions?

- **Initial Responder (10 min):** Michael Bratman
- **Second Responder (5 min):** Vincent Conitzer
- **Third Responder (5 min):** John-Dylan Haynes
- **Early-Career Responder (5 min):** Paul Talma
- **Chair:** Uri Maoz

Session 2: Responsibility, Law & AI Intentionality (11:15am~12:30pm)

Guiding Question: How should legal and ethical frameworks conceptualize intentionality as it pertains to AI? How can responsibility be assigned when AI actions (intended by the user or emergent) cause harm? Should concepts like 'mens rea' apply, or do they at a minimum require redefinition for artificial agents?

- **Initial Responder (10 min):** Scott Shapiro
- **Second Responder (5 min):** Uri Maoz
- **Third Responder (5 min):** Pamela Hieronymi
- **Early-Career Responder (5 min):** Ben Perry
- **Chair:** Gideon Yaffe

Afternoon Sessions

Session 3: Measuring, Modeling & Decoding Intentions (2:30pm~3:45pm)

Guiding Question: What empirical methods from neuroscience and psychology (e.g., neural decoding, behavioral analysis, disorder studies) can be adapted to measure, model, or infer intentions in AI? Conversely, how can AI models advance our understanding of human intentional processes?

- **Initial Responder (10 min):** John-Dylan Haynes
- **Second Responder (5 min):** Kyongsik Yun
- **Third Responder (5 min):** Bill Newsome
- **Early-Career Responder (5 min):** Alejandro de Miguel
- **Chair:** Patrick Haggard

AI
AI

Session 4: Biological vs. Artificial Intentions: A Comparative View (4:15pm~5:30pm)

Guiding Question: What can we learn by directly comparing concepts like purpose/function, goals, and intentions across diverse biological systems (shaped by evolution) and artificial systems (designed or learned)? What are the fundamental similarities and differences in their constraints, capabilities, and potential for goal development?

- **Initial Responder (10 min):** Tom Clandinin
- **Second Responder (5 min):** Colin Allen
- **Third Responder (5 min):** Hillard Kaplan
- **Early-Career Responder (5 min):** Dimitri Bredikhin
- **Chair:** Adina Roskies

Mini Session: June 2nd Summary (5:45pm~6:15pm)

Tomáš Dominik

AI
AI

June 3rd: Mechanisms & Interactions

Morning Sessions

Mini Session: June 3rd Look-Ahead (9am~9:30am)

Achintya Saha

Session 5: Mechanisms & Representation of Intention (9:30am~10:45am)

Guiding Question: Do intentions require explicit representation (neural, computational, symbolic)? How do mechanisms of intention formation, commitment, and execution differ between biological brains and AI architectures (e.g., RL, SSL), and what role, if any, do consciousness and intelligence play?

- **Initial Responder (10 min):** Michael Mozer
- **Second Responder (5 min):** Gideon Yaffe
- **Third Responder (5 min):** Gabriel Kreiman
- **Early-Career Responder (5 min):** Iwan Williams
- **Chair:** Vincent Conitzer

Session 6: Explainability, Opacity & Trust (11:15am~12:30pm)

Guiding Question: Given the inherent complexity and opacity of both advanced AI and human cognition, how can we develop reliable methods for explaining behavior and assessing the trustworthiness of stated intentions or hindsight rationalizations from either humans or AI systems?

- **Initial Responder (10 min):** Uri Maoz
- **Second Responder (5 min):** Adam Shai
- **Third Responder (5 min):** Adina Roskies
- **Early-Career Responder (5 min):** Lucas Jeay-Bizot
- **Chair:** William Newsome

Afternoon Sessions

Session 7: Alignment, Control & Predictability (2:30pm~3:45pm)

Guiding Question: What technical, architectural, and training methodologies are most promising for aligning complex AI behavior with human intentions and values, preventing unintended consequences or shortcut solutions? How can we manage risks, perhaps drawing parallels to human societal controls?

- **Initial Responder (10 min):** Vincent Conitzer
- **Second Responder (5 min):** Patrick Haggard
- **Third Responder (5 min):** Paul Riechers
- **Early-Career Responder (5 min):** Paulius Rimkevičius
- **Chair:** Gabriel Kreiman

AI
AI

Session 8: Intention in Action & Social Interaction (4:15pm~5:30pm)

Guiding Question: How do intentions structure planning, commitment, and action execution over time in humans and AI? Can AI effectively recognize, interpret, and participate in human social interactions involving individual and shared intentions (e.g., conversation, collaboration, games)?

- **Initial Responder (10 min):** Anna Leshinskaya
- **Second Responder (5 min):** Walter Sinnott-Armstrong
- **Third Responder (5 min):** Hillard Kaplan
- **Early-Career Responder (5 min):** Shaozhe Cheng
- **Chair:** Michael Mozer

Mini Session: June 3rd Summary (5:45pm~6:15pm)

Daniel Friedman

June 4th: Future Directions

Morning Session

Mini Session: June 4th Look-Ahead (9am~9:30am)

Ayana Shirai

Session 9: Emergent Goals, Agency & Conceptual Frameworks (9:30am~10:45am)

Guiding Question: What constitutes AI 'agency'? Under what conditions might AI develop genuinely novel goals or values? Are current folk psychological concepts adequate for understanding current and future AI, or will interaction with advanced AI reshape our own conceptual frameworks of mind and intention?

- **Initial Responder (10 min):** Walter Sinnott-Armstrong
- **Second Responder (5 min):** Sagi Perel
- **Third Responder (5 min):** Aaron Schurger
- **Early-Career Responder (5 min):** Lee Hristienko
- **Chair:** Pamela Hieronymi

AI
AI

Concluding Session

Session 10: Past & Future: Summary and Plans (11:15am~12:30pm)

What have we taken from the workshop?

Future plans: Workshop outcome (Book? Collection? Something else?)

Discussion of a hub for intentions & AI

Chair: Uri Maoz

WORKSHOP ATTENDEES

AI Researchers

Vincent Conitzer (Carnegie Mellon University)

Michael Mozer (Google DeepMind)

Sagi Perel (Google DeepMind)

Paul Riechers (Simplex AI Safety)

Adam Shai (Simplex AI Safety)

Kyongsik Yun (JPL/Caltech)

Philosophers/Legal Scholars

Colin Allen (UCSB)

Michael Bratman (Stanford University),

Pamela Hieronymi (UCLA)

Adina Roskies (UCSB)

Scott Shapiro (Yale University)

Walter Sinnott-Armstrong (Duke)

Gideon Yaffe (Yale University)

Neuroscientists/Biologists/Psychologists

Tom Clandinin (Stanford University)

Patrick Haggard (University College London)

John-Dylan Haynes (Charité-Universitätsmedizin Berlin)

Hillard Kaplan (Chapman University)

Gabriel Kreiman (Harvard University)

Anna Leshinskaya (UCI)

Uri Maoz (Chapman University)

William Newsome (Stanford University)

Aaron Schurger (Chapman University)

Early-Career Researchers

Dimitri Bredikhin (Chapman University)

Shaozhe Cheng (Duke University)

Alejandro de Miguel (Chapman University)

Tomáš Dominik (Chapman University)

Daniel Friedman (Stanford University)

Lee Hristienko (UCSB)

Lucas Jeay-Bizot (Chapman University)

Ben Perry (Texas Tech University)

Paulius Rimkevičius (Vilnius University)

Achintya Saha (Tennessee Tech University)

Ayana Shirai (Duke University)

Paul Talma (UCLA)

Iwan Williams (University of Copenhagen)

Agency, Intentions, and Artificial Intelligence