Deceptive & Counter-Deceptive Machines .5

(Founding) Symposium at IACAP 2013

When & Where: Specific T&P during conference span July 15–17 2013, TBA

Kubrick's landmark film *2001* features an at once deceptive and counter-deceptive machine (HAL 9000), and deceptive and (desperately) counter-deceptive humans. Is this volatile mixture our future in a microcosm? Yes, and the mixture is materializing before our very eyes, in no small part because: humans bent on doing great harm can only succeed if they deceive; our best bet for thwarting such humans is probably to enlist the power of counter-deceptive machines; deceptive machines are becoming a crucial part of the Defense arsenal, for PSYOPS and more; and so on.

But:

Can machines really deceive us? Can they deceive each other? What *is* deception? Can there be both a science and engineering of machine deception and counter-deception? If so, what would it look like? How can we have a science and engineering of trust in a machine-human space if we don't understand deception and counter-deception? How can we effectively use machines to counter deception perpetrated by machines, and by humans? …

On these topics, 4 presentations & a follow-on round-table discussion …
(titles, abstracts, bios follow)

A Future for Lying Machines

Micah Clark & **David Atkinson**

This talk addresses the present and future potential of autonomous systems that manipulate, mislead, and deceive. As we will show, such "lying machines" already exist, albeit in a nascent state. Lying machines have rather obvious applications in social networking, cyber-security, and state intelligence, as tools for both targeted subversion and broad persuasion campaigns. Interestingly, both lying machines and their dual — "trustworthy" AI — are made possible by an understanding of the human social interface for trust: one cannot technologically enable one without enabling the other. We argue that insofar as science is already reverse engineering the human interface for trust, we ought to consider strategically beneficial applications of deceptive as well as trustworthy AI, instead of pretending that lying machines will not one day emerge as a by-product.

Dr. **Micah Clark** is a Research Scientist at the Florida Institute for Human & Machine Cognition (IHMC), where he works on computational models of trust, belief, discourse, and theory of mind. Micah's research examines the development, maintenance, and manipulation of trust in the context of human conversations, human-machine interaction, and in predictive analytic tools and decision-support systems. Prior to joining IHMC, Micah spent 15 years at Caltech/NASA JPL working in the areas of simulation, fault management, and system autonomy for robotic explorers such as the Martian rovers, Spirit, Opportunity, and Curiosity. Micah received a PhD in Cognitive Science from Rensselaer Polytechnic Institute (RPI) in 2010 and a BS in Computer Science & Philosophy from RPI in 1999.

Dr. **David J. Atkinson**, currently Senior Research Scientist, Florida Institute for Human and Machine Cognition, is a computer scientist with career-long experience in research, development, and oversight positions addressing the full lifecycle of intelligent, autonomous systems. Atkinsons current research focuses on human-machine trust and architectures for intelligent, autonomous systems. Atkinson received the Doctor of Technology (D. Tech.) in Computer Systems Engineering from Chalmers University of Sweden in Göteborg, Sweden, and was named a Docent of the University. He was awarded the Master of Science (MS) and Master of Philosophy (MPhil) degrees in Computer Science from Yale University and the Bachelor of Arts (BA) in Psychology from the University of Michigan. Atkinson supported the Air Force Office of Scientific Research (AFOSR) where he was a Program Manager in the Asian Office of Aerospace R&D. As additional duty, he was Program Manager for the Robust Computational Intelligence program at AFOSR. Prior to joining IHMC, he worked at Caltech/JPL. Atkinson received the NASA Exceptional Service Medal (1990) and other awards for achievements in AI and robotics. Dr. Atkinson has over 40 peer-reviewed publications and numerous other technical reports in the areas of artificial intelligence, aerospace systems, spaceflight operations, and robotics

Mindreading Deception in Dialog

Will Bridewell & **Alistair M.C. Isaac**

This talk considers the problem of detecting deceptive agents in a conversational context. We argue that distinguishing between types of deception is required to generate successful action. This consideration motivates a novel taxonomy of deceptive and ignorant mental states, emphasizing the importance of an ulterior motive when classifying deceptive agents. After illustrating this taxonomy with a sequence of examples, we introduce a Framework for Identifying Deceptive Entities (FIDE) and demonstrate that FIDE has the representational power to distinguish between the members of our taxonomy. We conclude with some conjectures about how FIDE could be used for inference.

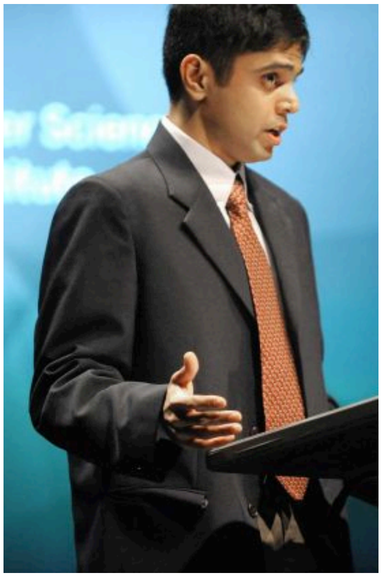
Will Bridewell is a research scientist at the Stanford Center for Biomedical Informatics Research (BMIR). He earned his PhD in Computer Science in 2004 from the University of Pittsburgh, where he developed a simple method for detecting negation in medical records and a unique approach to explaining anomalies in scientific data. Afterwards, he moved to the Computational Learning Laboratory at Stanford University, where he furthered research in inductive process modeling and other forms of computational scientific discovery. He joined BMIR in 2009 to develop a new approach to abductive inference that would form the foundations of a system that supports socially aware inference. In May 2013, he will join the Naval Center for Applied Research in Artificial Intelligence at the Naval Research Laboratory.

Alistair Isaac earned his PhD in Philosophy from Stanford University in 2010. After post-docs at the University of Michigan and the University of Pennsylvania, he will start a Lectureship at the University of Edinburgh in August, 2013. His work focuses primarily on philosophy of psychology and cognitive science, with a special interest in topics relating to perception and bounded rationality.

Games for Deception, Counter-Deception & Cyberwarfare (**DCC**)

Naveen Sundar Govindarajulu

Is it possible to harness a gaming crowd for the trio of **DCC** with playable computer games? A survey of the state of the art in using games for **DCC** will be presented. A vocabulary and criteria for designing and talking about different games that could be used in the wide space of not only training experts, but also recruiting the public in posssible large-scale **DCC** will be discussed. The talk will include a preview of X, a new game for crowd sourcing in this space. The talk ends with the posing some new ethical questions that could arise due to such games.



Naveen Sundar Govindarajulu is a PhD candidate in Computer Science at RPI. His dissertation centers around designing and implementing the world's first *uncomputable games*, and has been partly funded by the International Fulbright Science and Technology Award. Before RPI, Naveen studied physics and electrical engineering (MSc & BE), and worked on applying statistical machine learning to biometrics and shape recognition.

Toward the Automated Detection of Deceptive Valuation

Alexander Bringsjord

The now-infamous acquisition by Hewlett Packard of former-UK-software-maker, Autonomy, in October of 2011, has been the largest buyout by HP to date; the price tag: a rather non-trivial $11 billion. A stance is not taken herein as to whether or not it was Autonomy or HP at fault for a valuation inflated by more than $3 billion; only time and (lengthy) litigation will tell ground truth. But the fact remains that the valuation *was* stunningly too high, and that accounting errors occurred. (Actual expense accounts were incorrectly associated with actual dollar amounts, thus increasing Autonomy's bottom line.) In an effort to demonstrate how M&A models, relevant agents, fabricated events, etc. can be formally represented, and thus, in turn, how detection of fraud involving such representations can be automated, we hypothetically assume, solely for scientific and engineering purposes, that agents on an "Autonomy side" of a mega-acquisition intentionally, and therefore fraudulently, inflate their valuation. We assume further that (human) agents on behalf of this side form beliefs about the beliefs held by finance folks on the "HP side" regarding what should appear in the financial statements to justify a buyout of $11 billion — which means that fraud and counter-fraud in the M&A realm involve what has been called "mindreading."

Alexander Bringsjord is Co-Founder (2009), President, and CEO of Motalen, Inc., a digital media company focused in the mobile space, and committed to producing content that not only entertains, but also develops, expands, and harnesses the human intellect. He holds the BS *cum laude* in Business Management & Philosophy, and will receive the MS in Accounting from the University of Albany in 2014, by which time he is slated to be a Certified Fraud Examiner. Alexander is the author of numerous publications in the intersection of business, computation, philosophy, and the mind, and actively speaks and consults in the areas of counter-fraud/accounting, entrepreneurship, and the philosophical foundations of business and economics.