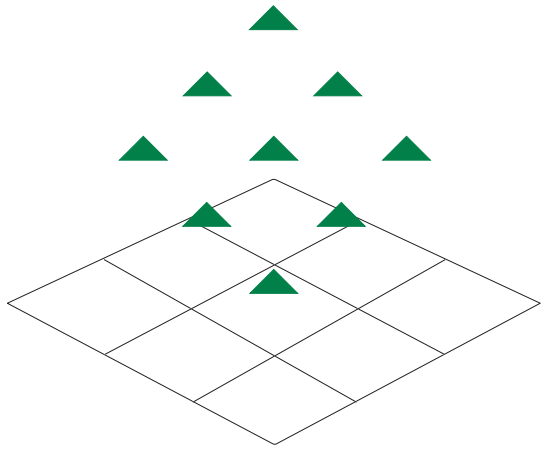
Reports on the 2018 AAAI Spring Symposium Series

Christopher Amato, Haitham Bou Ammar, Elizabeth Churchill, Erez Karpas, Takashi Kido, Mike Kuniavsky, W. F. Lawless, Frans A. Oliehoek, Francesca Rossi, Stephen Russell, Siddharth Srivastava, Keiki Takadama, Philip van Allen, K. Brent Venable, Karl Tuyls, Peter Vrancx, Shiqi Zhang

■ *The Association for the Advancement of Artificial Intelligence, in cooperation with Stanford University's Department of Computer Science, presented the 2018 Spring Symposium Series, held March 26–28, 2018, on the campus of Stanford University. The seven symposia held were AI and Society: Ethics, Safety, and Trustworthiness in Intelligent Agents; Artificial Intelligence for the Internet of Everything; Beyond Machine Intelligence: Understanding Cognitive Bias and Humanity for Well-Being AI; Data-Efficient Reinforcement Learning; The Design of the User Experience for Artificial Intelligence (the UX of AI); Integrated Representation, Reasoning, Learning, and Execution for Goal-Directed Autonomy; Learning, Inference, and Control of Multiagent Systems. This report, compiled from organizers of the symposia, summarizes the research of the symposia that took place.*

AI and Society: Ethics, Safety, and Trustworthiness in Intelligent Agents

Artificial intelligence has become a major player in today's society and that has inevitably generated a proliferation of thoughts and sentiments on several related issues. Many, for example, have felt the need to voice, in different ways and through different channels, their concerns on the possible undesirable outcomes caused by artificial agents, the morality of their use in specific, sensitive sectors, such as the military, and the impact these agents will have on the labor market. The AAAI Spring 2018 symposium, AI and Society: Ethics, Safety, and Trustworthiness in Intelligent Agents, succeeded both in gathering a diverse group of researchers from



many disciplines and in fostering a scientific discussion on this topic. Joining AI researchers were philosophers, economists, sociologists, and representatives of industry for what proved to be a fruitful and stimulating conversation.

The symposium successfully attracted contributions on a broad set of topics related to the ethics, safety, and trustworthiness of AI. The conversation and presentations focused on the adoption of a scientific approach to help understand more fully the impact of AI and to put into perspective the multitude of opinions on these matters. We received submissions from different disciplines and with different perspectives, some focusing on specific and technical details, others bringing a more general point of view to the desiderata for AI in terms of society. Some papers also addressed both short-term and long-term analysis of AI's impact on different aspects of society.

The symposium included two invited talks. The first, by Vince Conitzer (Duke University), presented recent work and several points for further discussion on moral artificial intelligence, kidney exchanges, and societal trade-offs. The second invited talk, by Emma Brunskill (Stanford University), addressed reinforcement learning in high-stakes domains. The symposium included six technical sessions on moral decision-making, ethics and moral agents, beneficial AI and AI divide, trustworthiness, ethics and value alignment, and, finally, applications and interactive agents. Also two stimulating discussion sessions were led by the Venerable Tenzin Priyadashi (MIT Media Lab), with Judy Wajcman (London School of Economics) and Francesca Rossi (IBM Research and University of Padova). Several key points were addressed during the talks and further elaborated on in the discussion sessions, including ethics in game theory, subjective and ethical preferences, morality in bottom-up ML frameworks, the role of creativity to solve ethical dilemmas, the meaning of contextual knowledge in AI ethics, research versus deployment regulations, and the importance of the concept of human well-being.

The symposium was organized by Francesca Rossi (IBM Research and University of Padova), K. Brent Venable (Tulane University and IHMC), and Toby Walsh (Data61, UNSW and TU Berlin). Rossi and Venable prepared this report. The papers of the symposium were published in the AAAI digital library.

Artificial Intelligence for the Internet of Everything

The Internet of Everything (IoE) generalizes machine-to-machine communication for the Internet of Things (IoT) to encompass people, robots, machines, and teams. At this symposium, we learned that IoE may revolutionize the way we humans do business, the way we communicate, create jobs, and govern, the way we educate and care for ourselves, all while helping us to make better decisions, be more productive, and innovate faster.

IoT is about connecting a network of static and mobile objects to enable them and humans to collect and share data. With the approach of IoT in everyday life, on battlefields (IoBT), in medicine (IoMT), industry (IIoT), and with intelligent devices (IoIT), some of the known issues are the explosion of data (for example, cross-compatible systems, storage locations); security (for example, password authentication, data exfiltration, covert channels, privacy); and risks to users, teams, enterprises, and institutions. IoE may be automatic or autonomous. It will likely manifest as heterogeneous and self-organizing complex systems that define human and machine processes, requiring interoperability, just-in-time (JIT) interaction, and the orchestration of local-adaptation functions to achieve objectives and goals.

There are also practical considerations to take into account. Whatever the systems, in daily use, each must be robust to interruption, failure, and wear and tear. Systems must have manual control backups and access to power (for example, robotic vacuum cleaners that recharge autonomously); user-friendly methods for joining and leaving networks; autonomous software updates and backups; and autonomous hardware updates (for example, ordering parts automatically or autonomously responding to recalls). A system must also provide forensic evidence in the event of a mishap with onboard and online recorders.

Open questions remain. Will systems communicate with each other or be independent? Will humans always need to be in the loop? Will systems communicate only with humans, or also with robots and machines? What are the policy and organizational implications of thing-based systems?

Linking this symposium with our 2016 AAAI symposium on using AI to reduce human error, intelligence may be critical to overcoming barriers when IoE addresses safety. For example, a fighter plane can already take control to save itself if its pilot loses consciousness during a high-g maneuver. If monitoring IoE systems with AI and team metrics, Germanwings flight 9525 might have safely secured itself by isolating the suicidal copilot, thus saving the lives of all aboard that day. Similarly, the speeding Amtrak train that derailed in 2015 as its head engineer lost awareness could have been spared the loss of life had the train taken control until it had safely stopped.

At the symposium, we discussed the recent lethal accident with a self-driving car in Arizona and the potentially lethal act of a swarm of drones sent against a Russian base in Syria. But as IoE evolves, when a machine harms a human, we need to know its decision process in its determinable context.

Neil Gershenfeld, director of MIT's Center for Bits and Atoms, one of our invited speakers, discussed

Photo courtesy, iStock

what may happen when "things" begin to think and reproduce as digital and physical worlds merge. Our other invited speakers addressed policy issues (Barry Horowitz, UVA); multiagent inference (Georgiy Levchuk, Aptima); intelligence with battlefield things (Alexander Kott, ARL); machines as teammates (Joseph Lyons, AFRL); dynamic agent management (Hesham Foaud, NRL); IoBT complexity (Steve Russell, ARL); and the economics of things (Shu Heng-Chen, Taiwan). Regular speakers added ecosystem models, compositional models, smart entities, distributed ledgers, message pipelines, interdepend-

ent teams, agent learning, smart layers, and valuable information.

The current landscape of IoE is characterized by a dramatic increase in scale and complexity across multiple dimensions, not just limited to technological capability. The pervasiveness of things and the subsequent benefits will dramatically change all aspects of human activity, from industrial to medical, from combat to philosophy. IoE will help society to evolve when "things" and humans are able to team together, as a collective intelligence, to help each other determine contexts, solve problems, reduce errors, and save lives.

The symposium was organized by Ranjeev Mittu, Donald Sofge, and Ira S. Moskowitz (Naval Research Laboratory), Stephen Russell (Army Research Laboratory), and W. F. Lawless (Paine College). This report was prepared by Russell and Lawless. The papers of the symposium were published in the AAAI digital library.

Beyond Machine Intelligence: Understanding Cognitive Bias and Humanity for Well-Being AI

In this AAAI spring symposium, we discussed cognitive bias and humanity in the context of well-being AI. We defined "well-being AI" as an AI research paradigm for promoting psychological well-being and maximizing human potential. The goals of well-being AI are (1) to understand how our digital experience affects our health and our quality of life and (2) to design well-being systems that put humans at the center. The important challenges of this research are how to quantify subjective things such as happiness, personal impressions, and personal values, and how to transform them into scientific representations with corresponding computational methods.

One of the important touchstones in understanding machine intelligence as it relates to human health and wellness is cognitive bias. Advances in big data and machine learning should not overlook some new threats to enlightened thought, such as the recent trend of using social media platforms and commercial recommendation systems to manipulate people's inherent cognitive biases.

The second important touchstone is humanity. As machine learning gains ground, rational thinking, on which early AI researchers had been focused, is rapidly replacing human thinking. Many might have begun to believe that irrational thinking itself constituted the root of humanity. Several discussions centered on the relationship of AI to humanity, both empirically and philosophically.

Our symposium included four invited talks to provide new perspectives on the limitations of current machine intelligence and the challenges for understanding humanity. Pang Wei Koh (Stanford University) gave a talk on understanding black-box deep learning predictions with influence functions. Avanti Shrikumar (Stanford University) discussed the issues of interpretable deep learning for genomics. Robert Reynolds (Wayne University) introduced the game-theoretic knowledge distribution in cultural algorithms. Finally, Daniel Martin (California State University) introduced research topics on human compassion, including stress reduction skills and competency development related to social capital and AI from the perspective of social psychology.

The symposium included 20 papers and three posters and demonstrations, presented over the course of the two and a half days. Topics included cognitive bias and humanity; understanding machine and human; measuring health; better health; understanding human: dementia; understanding society: decision support; and measuring performance.

Among the papers presented, Takashi Kido (Preferred Networks) discussed the challenges in understanding cognitive bias and humanity for well-being AI. Christina Alexandris (National University of Athens) discussed the issues of measuring cognitive bias in spoken interaction and conversation and of generating visual representations. Keiki Takadama (The University of Electro-Communications) introduced research on providing understandable knowledge (correct and commonly accepted knowledge) with machine learning in the care support domain. Ayae Ide (National Institute of Advanced Industrial Science and Technology) proposed a policy-decision support system for an aging society based on probabilistic latent spatial semantic structure modeling. Finally, Sachiko Deguchi (Kindai University) introduced a study on the UI and musical performance system and score representation.

The symposium provided researchers with diverse backgrounds unique opportunities for coming together and incubating new ideas through innovative and constructive discussions. The material presented explored important interdisciplinary challenges for guiding future advances in the AI community.

Takashi Kido and Keiki Takadama served as cochairs of this symposium. The papers of the symposium were published in the AAAI digital library.

Data-Efficient Reinforcement Learning

Sequential decision-making (SDM) is an essential component for autonomous systems. Although significant progress has been made towards developing algorithms for solving isolated SDM tasks, these algorithms often require large amounts of experience before achieving acceptable performance. Unfortunately, interactions in real-world environments can be costly, especially when initial performance is poor. Solving important real-world problems often requires

a well-defined, acceptable baseline and highly sample-efficient learning. This is particularly true for high-dimensional tasks, such as robotics control or general game-playing environments. Multiple methods have been proposed for efficient reinforcement learning algorithms that can generalize well to unobserved environments or situations. This symposium brought together researchers from reinforcement learning, probabilistic modeling, robotics, and multiagent systems to discuss potential solutions to these challenging problems. The symposium featured a range of paper presentations and invited talks, with an emphasis on theoretically grounded approaches for data-efficient learning.

A major theme of papers presented at the symposium was the use of mathematical frameworks from probabilistic modeling and information theory in reinforcement learning. The first day started with an invited talk by Prof. Warren Powell, who introduced a unified mathematical framework for stochastic optimisation. This talk was followed by several contributed papers that discussed probabilistic methods to deal with uncertainty in reinforcement learning problems.

A second track that was strongly represented was the use of models and simulated data for training reinforcement learning agents. The use of simulation was a major theme in the invited contribution by Josiah Hanna (University of Texas at Austin). This talk presented an overview of related ongoing research projects at Peter Stone's Learning Agents Research Group. Several papers then discussed the use of different model-based techniques to improve the sample efficiency of basic reinforcement learning. The symposium concluded with a final invited talk by Sofia Ceppi (PROWLER.io). Ceppi discussed a framework for combining game theory and mechanism design with agents learning from interaction.

The main takeaway message of the symposium was the general need to develop theoretically grounded approaches that use probabilistic modeling as a foundation to learn and make decisions in real-world environments. Participants discussed the strong focus on black-box deep learning models underlying current trends in AI. While these methods have undoubtedly enabled learning to scale up in sequential decision problems, more work is needed to make learning agents deal with uncertainty in a principled way.

Dongho Kim (PROWLER.io CTO) chaired the symposium. Haitham Bou Ammar and Peter Vrancx prepared this report.The papers of the symposium were published in the AAAI digital library.

The Design of the User Experience for Artificial Intelligence

As AI is rapidly becoming part of everyday consumer and professional systems, designers must adapt to a domain with sometimes radical new requirements, technologies, affordances, and constraints compared to other design contexts. Most importantly, AI introduces new interactions between people and AI, and between AI and other AI systems. The goal of this symposium was to bring together a diverse group of practitioners and researchers, creating opportunities for rich cross-fertilization and building a shared understanding of the challenges and opportunities in AI design.

The symposium participants ranged from designers who are working on real-world products such as Spotify or mission-critical applications such as NASA ISS procedure automation, to researchers investigating the effects of algorithmic transparency on user perceptions, to graduate students creating speculative AI projects, to AI theoreticians proposing new design goals, to tool developers creating new ways of prototyping AI.

In addition to paper presentations and posters, there were two moderated discussions: one on AI design tool needs, and one on explainable AI design issues. There was also an expert panel on cybernetics and design, with design planner and teacher Hugh Dubberly (Dubberly Design Office), primatologist Deborah Forster (Contextual Robotics Institute, UC San Diego), and UX designer Jody Medich (Singularity University Labs). The panel was moderated by interaction design researcher Wendy Ju (Cornell Tech) and mechanical engineer Nik Martello (PhD student, Stanford).

The clear challenge was resisting the temptation to overdefine the boundaries of either user experience design or AI. We acknowledged they're both expansive umbrella fields that encompass many subdisciplines from human computer interaction and product strategy in the case of user experiences, to computer vision and process planning in AI. The cochairs focused the symposium on interesting spaces in the intersection of these disciplines, whatever the boundaries, on approaches and challenges for designing experiences for current AI systems, and on the design implications of future systems.

Out of this vast discussion space developed a strong need for shared language and developing an understanding of AI as a medium. We discussed how AI algorithms, data, and training all combine to form the material to be designed with. We also discussed the need for a richer cooperation between design and the engineering and science of AI, so that designers, in addition to benefiting from new capabilities, can positively influence the development of future AI based on their understanding of human applications and requirements. Many other themes emerged in this complex domain. Some key issues follow.

AI collaborators: Rather than seeing AI as a way to automate activities or provide solutions, AI systems can be designed as collaborators that participate with humans in creating shared outcomes. Such an

approach would take into account concepts around cybernetics, distributed cognition, the limits of narrow AI, and the complexity of human creativity.

Speculative design: Designers have a long history of practice that is oriented not towards solving a problem, but rather exploring a new domain or future potential. Design speculation for AI can consider the new opportunities, the ethical, cultural, and societal impacts, and the potential hazards.

Cybernetics: As designers move from the design of individual things to the complex ecologies of smart things, cybernetics has renewed relevance for designers in terms of understanding the dynamics of systems, goals, feedback, and conversations.

Design tools: Given the differences in design goals and strategies for autonomous systems, new tools are needed that help designers build working prototypes for both exploration and application. These tools should provide ways of working around the difficult aspects of AI and provide ways for designers and others to quickly experiment and iterate so they can build understanding and design better systems.

Explainable AI: Beyond the technical challenges of XAI, our discussion focused on the design issues involved. What affordances can we make available so the user can respond to explanations? How much explanation should there be, and how much is too much? What is the role of trust? What if an AI decision is nonintuitive? Does the public need to understand how AI makes decisions?

Elizabeth Churchill (Google), Mike Kuniavsky (Xerox Parc), and Philip Van Allen (Art Center College of Design) served as the cochairs of this symposium, with the help of Molly Steensen (Carnegie Mellon University). The papers of the symposium were published in the AAAI digital library.

Integrating Representation, Reasoning, Learning, and Execution for Goal-Directed Autonomy

Recent advances in AI and robotics have led to a resurgence of interest in the objective of producing intelligent agents that help us in our daily lives. Such agents must be able to rapidly adapt to the changing goals of their users, and the changing environments in which they operate. These requirements lead to a balancing act that most current systems have difficulty contending with: on the one hand, human interaction and computational scalability favor the use of abstracted models of problems and environment domains; on the other, generating goal-directed behavior in the real world typically requires accurate models that are difficult to obtain and computationally hard to reason with.

This symposium addressed the core research questions that arise in designing autonomous systems that execute their actions in complex environments using imprecise models. The sources of imprecision may range from computational pragmatism to imperfect knowledge of the actual problem domain.

The symposium brought together researchers from a variety of subfields of AI such as robot planning, model error detection, reasoning with abstractions, statistical learning for sequential decision-making and robotics, and cognitive systems. The symposium featured presentations of 25 accepted papers in addition to the invited talks. These presentations included short, position-paper presentations as well as longer presentations for full technical papers. The audience participated actively in the presentations using allocated discussion times in each presentation session. The symposium also hosted three invited speakers: Jeremy Frank (NASA Ames Research Center), David Aha (US Naval Research Laboratory), and Emma Brunskill (Stanford University). Finally, the attendees visited the Stanford Robotics Lab, where they were hosted by Oussama Khatib and Mikael Jorda, who explained the OceanOne robot and demonstrated haptic control devices.

One of the main themes of the symposium was the notion of discrepancies, particularly discrepancies between the expected state of the world according to a model and the observed state of the world. Such discrepancies can be used to trigger a correction to the model or a refinement of the abstraction used in creating the model. They could also be used to trigger goal reasoning, as they might imply that the goal currently being pursued by the system is irrelevant, or that there are more important goals to pursue at the moment.

Siddharth Srivastava, Shiqi Zhang, Nick Hawes, Erez Karpas, George Konidaris, Matteo Leonetti, Mohan Sridharan, and Jeremy Wyatt served as cochairs of this symposium. Siddharth Srivastava, Shiqi Zhang, and Eraz Karpas prepared this report. The papers of the symposium were published in the AAAI digital library.

Learning, Inference, and Control of MultiAgent Systems

Agents are and will be deployed in a range of environments. They will need to compete in market places, to cooperate in teams, to communicate with others, to coordinate their plans, and to negotiate outcomes. Examples include self-driving cars interacting in traffic, personal assistants acting on behalf of humans and negotiating with other agents, swarms of unmanned aerial vehicles, financial trading systems, robotic teams, and household robots. Multiagent systems can have desirable properties such as robustness and scalability, but their design

requires careful consideration of incentive structures, learning, and communication.

The symposium brought together researchers from different fields to discuss the current state of multiagent learning, as well as future directions and roadblocks. The main topics of discussion concentrated on questions such as the following. What are the right models for multiagent learning in different situations (for example, Dec-POMDPs, I-POMDPs, other decision-theoretic or game-theoretic models)? What are the best benchmarks to use and how can we create new, high-quality ones? What is the role of deep learning in multiagent learning? What are the best metrics for evaluating multiagent learning performance? How do we ensure that multiagent learning is applicable to real-world problems?

The invited speakers spoke on many of these issues. Ann Nowé (Vrije Universiteit Brussel) discussed learning properties with simple learning rules such as learning automata. Igor Mordatch (OpenAI) talked about recent methods for deep reinforcement learning for communication in cooperative multiagent systems. Mac Schwager (Stanford University) talked about multiagent learning for multirobot coordination. Mykel Kochenderfer (Stanford University) spoke about multiagent learning for applications ranging from aircraft collision avoidance to autonomous vehicles and drones. Pradeep Varakantham (Singapore Management University) described combinations of game theory and optimization in order to balance resource demand, and Emma Brunskill (Stanford University), in a joint session with the Symposium on Integrated Representation, Reasoning, and Learning in Robotics, talked about progress in model-based reinforcement learning by using ensembles of neural networks as models.

Ten contributed talks were also given on many of the topics previously mentioned. These talks described work on new game-theoretic and decision-theoretic methods in scenarios that are partially observable, human-interactive, ad-hoc, imperfect information, multi-task or nonstationary. The scope of work represents the broad set of approaches and situations in which multiagent learning applies.

Overall, there was much enthusiasm about the future of multiagent learning. A number of topics relevant for the progress of the field, such as scalability, evaluation, and properties that are relevant for real-world applications, were deliberated in a discussion session and analyzed in further detail in breakout sessions. One point that resonated particularly well with the audience was the idea of constructing a suite of benchmark problems for multiagent learning. It was agreed that future symposia and workshops will be held to continue discussion and the progress that has been made. Concrete goals include an industry and academe partnership to develop a website with papers and benchmarks, as well as widening the participation in the discussion by including additional senior and junior researchers.

This symposium was organized by Chris Amato (Northeastern University), Thore Graepel (DeepMind), Joel Leibo (DeepMind), Frans Oliehoek (Delft University of Technology and University of Liverpool), and Karl Tuyls (DeepMind). Christopher Amato, Frans Oliehoek, and Karl Tuyls prepared this report. The papers of the symposium were published in the AAAI digital library.

Christopher Amato is an assistant professor at Northeastern University.

Haitham Bou Ammar is the head of reinforcement learning at PROWLER.io, Cambridge, UK.

Elizabeth Churchill is the director of UX at Google.

Erez Karpas is a senior lecturer at Technion, the Israel Institute of Technology.

Takashi Kido is a researcher in Japan. He had been a visiting researcher at Stanford University.

Mike Kuniavsky is a user experience designer at Parc.

W. F. Lawless is a professor at Paine College.

Frans A. Oliehoek is an associate professor at TU Delft and the University of Liverpool.

Francesca Rossi is a distinguished research scientist at the IBM T. J. Watson Research Center, and professor of computer science at the University of Padova, Italy.

Stephen Russell is a researcher at the US Army Research Laboratory.

Siddharth Srivastava is an assistant professor at Arizona State University.

Keiki Takadama is a professor of the University of Electro-Communications in Japan.

Karl Tuyls is a research scientist at Google DeepMind.

Philip Van Allen is a professor in the Media Design Practices department at the Art Center College of Design.

K. Brent Venable is a professor of computer science at Tulane University and a research scientist at the Florida Institute for Human and Machine Cognition (IHMC).

Peter Vrancx is a senior machine learning researcher at PROWLER.io, Cambridge, UK.

Shiqi Zhang is an assistant professor at Cleveland State University.