Reports of the Workshops Held at the 2019 AAAI Conference on Artificial Intelligence

Guy Barash, Mauricio Castillo-Effen, Niyati Chhaya, Peter Clark, Huáscar Espinoza, Eitan Farchi, Christopher Geib, Odd Erik Gundersen, Seán Ó hÉigeartaig, José Hernández-Orallo, Chiori Hori, Xiaowei Huang, Kokil Jaidka, Pavan Kapanipathi, Sarah Keren, Seokhwan Kim, Marc Lanctot, Danny Lange, David Martinez, Marwan Mattar, Mausam, Julian McAuley, Martin Michalowski, Reuth Mirsky, Roozbeh Mottaghi, Joseph C. Osborn, Julien Pérolat, Martin Schmid, Arash Shaban-Nejad, Onn Shehory, Biplav Srivastava, William Streilein, Kartik Talamadupula, Julian Togelius, Koichiro Yoshino, Quanshi Zhang, Imed Zitouni

■ *The workshop program of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) was held in Honolulu, Hawaii, on Sunday and Monday, January 27 and 28, 2019. There were 16 workshops in the program: Affective Content Analysis: Modeling Affect-in-Action; Agile Robotics for Industrial Automation Competition; Artificial Intelligence for Cyber Security; Artificial Intelligence Safety; Dialog System Technology Challenge; Engineering Dependable and Secure Machine Learning Systems; Games and Simulations for Artificial Intelligence; Health Intelligence; Knowledge Extraction from Games; Network Interpretability for Deep Learning; Plan, Activity, and Intent Recognition; Reasoning and Learning for Human-Machine Dialogues; Reasoning for Complex Question Answering; Recommender Systems Meet Natural Language Processing; Reinforcement Learning in Games; and Reproducible AI. This report contains brief summaries of all the workshops that were held.*

Affective Content Analysis and CL-Aff Shared Task: In Pursuit of Happiness

The Affective Content Analysis workshop series held at the AAAI Conference on Artificial Intelligence is an interdisciplinary platform intended to engage the AI and machine-learning communities about open problems in affective content analysis and understanding, with a special focus on affect in language and text. The theme of this second workshop was modeling affect-in-action, with a shared task (CL-Aff — in pursuit of happiness) to encourage the development of new models and approaches for modeling happy moments.

ISSN 0738-4602

Affective computing has traditionally focused on modeling human reactions using multimodal sensor data but not using text. Sentiment and emotion analysis, on the other hand, has been applied on text as well as multimodal data sets, but this research has been limited to quantifying well-defined human reactions. Affect analysis, that is, techniques and applications that understand the experience of an emotion in the context of language and text, is an upcoming research space. Little has been done to explore the affective facets of dynamic or multimedia data. Furthermore, the subjective nature of human affect suggests the need to measure in ways that recognize multiple interpretations of human responses. Other challenges include standardizing the measurement of affect to meaningfully compare different affective models against each other, addressing the challenges in cross-media, cross-domain, and cross-platform affect analysis, and identifying consumer psychology theories and behaviors related to affect, which are amenable to computational modeling

The workshop program focused on the analysis of emotions, sentiments, and attitudes in textual, visual, and multimodal content for applications in psychology, language understanding, and computer vision. Besides original research presentations and posters, the workshop hosted a range of keynote speakers, whose presentations highlighted the state of the art in affective computing in a range of fields.

Alon Halevy talked about various efforts toward taking affect analysis techniques to practice, specifically focusing on affective search. Rada Mihalcea from the University of Michigan discussed the importance of grounding emotion and affect analysis in context. The study emphasized the importance of looking beyond the language, including user preferences and environmental variables. Ellen Riloff from the University of Utah shared an analysis of affective events and reasons behind their polarity. Atanu Sinha from Adobe Research shared two studies that look at affect understanding and interpretation from a consumer psychology perspective. The talk focused on studying facial expressions and their impact on offers and counteroffers in a negotiation context. Finally, Lyle Ungar from the University of Pennsylvania talked about studying empathy in social media content. He introduced the challenges in building computational models for psychology-based theories and discussed novel methodologies based on the machine-learning model

The workshop concluded with a panel discussion among the keynote speakers, moderated by the organizers, on potential directions for future events and the scope of interdisciplinary research. The papers of the workshop were published as *CEUR Workshop Proceedings*, Volume 2328. The workshop was cochaired by Niyati Chhaya and Kokil Jaidka, who also wrote this report.

Artificial Intelligence for Cyber Security Workshop

The 2019 Artificial Intelligence for Cyber Security workshop focused on research and applications of AI to operational problems in cybersecurity, including machine learning, game theory, threat modeling, and automated and assistive reasoning. The workshop began with a keynote speech by Craig Knoblock, executive director of USC/ISI, on building knowledge graphs for cybersecurity. Knoblock began with a brief overview of graph analysis, including a useful analysis pipeline to frame the application of graphical techniques to complex data. He provided several motivating examples, including recent application of the techniques to the problem of forecasting evolving cyber threats within an enterprise environment.

The initial session featured several talks on the application of AI to problems in cybersecurity. The first paper presented results on leveraging Markov game modeling of moving target defenses to protect against multistage cyber threats in a cloud network environment. Results compared favorably with static nongraph-based techniques. The second paper presented work in leveraging game-theoretical models to optimize defenses for so-called watering hole attacks. Experimental analyses demonstrated the benefit of the approach. A third paper explored the application of planning and model-based diagnosis to automate the process of cyber physical system design while at the same time meeting security constraints. The technique's effectiveness was evaluated on an autopilot model.

The first three workshop papers were followed by a panel discussion: Thirsty in the Age of Plenty — A Discussion on (Lack of) Datasets for AI and Cyber Security. The panel participants were from industry, government, and academia. The IMPACT data repository, hosted by the Department of Homeland Security, contains a corpora of network data, cyber defense data, and cyber attack data. Two important takeaways from the panel were the need for a definition of success when defending networks and the need for more representative data, such as data from realistic live environments. It was also pointed out by the panelists that data used in operational environments are different from data used for advancing research. The data for research are best when these data represent evolving threats.

The morning session concluded with two technical papers. The first presented work in using novel data masking techniques to protect a machine-learning classifier. Theoretical guarantees of the technique were developed and evaluated on benchmark data sets. The final paper of the morning discussed the use of fuzzy hashes extracted from kernel embeddings to enable file matching, despite adversary insertion and deletion operations. Results compare quite favorably with standard hashing techniques used to recognize malware.

The afternoon keynote presented by Una-May O'Reilly, professor and research scientist at MIT/CSAIL, dealt with artificial adversary intelligence. O'Reilly began her keynote with a review of recent work in generating adversarial malware examples while still preserving malicious function. Capturing the conflicting defender goal of minimizing error and the attacker goal of maximizing misclassification, results showed enhanced robustness for machine learning algorithms trained on generated adversarial examples. O'Reilly then described work in applying principles of coevolution to understanding and predicting behavior of cyber attackers and defenders. Leveraging a game theoretical representation, results were presented showing the technique can be used to anticipate attacks to better protect resources.

Following the afternoon keynote, a presentation focused on adversarial attacks of speech and text. Adversarial examples are crafted by adding human-imperceptible perturbations to inputs such that a machine-learning-based classifier incorrectly labels them. Inspired by multiversion programming, the authors proposed a novel audio detection approach that utilizes multiple off-the-shelf automatic speech recognition systems to determine whether an audio input is an adversarial example. The approach is based on identifying malware and distributing the detection across multiple programs in parallel to achieve expedience. The evaluation shows that the detection achieves accuracies of greater than 98.6 percent. The next paper addressed preventing adversarial attacks against networks. The assumption is that attackers typically just add to existing software to ascertain that application programming interfaces still work, therefore avoiding detection. The authors were able to demonstrate fairly good performance with an area under the curve in the high 90s. The last paper of the workshop, before the discussion and the presentation of the winning paper of the AICS challenge problem, was on generating adversarial samples. The author's approach was to not disturb the image but minimally modify it while preserving the image's label. The technique relied on performing a linear combination of power functions while maintaining a structure-preserving transformation.

The final session of the workshop was devoted to the AICS 2019 challenge, which posed the problem of correctly classifying malware types while under an adversarial learning attack. After a brief overview of the challenge problem motivation and creation process, by Jason Matterer, the authors of the winning submission presented their approach to the problem. The winning approach leveraged an ensemble of classifiers, trained on transformed training data to reduce the impact of evasion techniques, resulting in the best overall score among the competing submissions. The authors codified their winning approach in a series of principles and a general framework that can be used by others to address adversarial learning attacks in general.

The workshop was cochaired by William Streilein, David Martinez, Jason Matterer, Una-May O'Reilly, Howie Shrobe, and Arunesh Sinha. The papers of the workshop were published on arXiv. This report was written by William Streilein and David Martinez.

Artificial Intelligence Safety

Safety in AI should be not an option but a design principle. However, there are varying levels of safety, diverse sets of ethical standards and values, and varying degrees of liability, all of which force us to consider trade-offs and multiple solution alternatives. These choices can only be analyzed holistically by integrating the technological and ethical perspectives into the engineering problem and by considering both the theoretical and practical challenges for AI safety. The AAAI-19 Artificial Intelligence Safety workshop explored these issues through a wide range of AI paradigms, considering systems that are application specific and those that are general. The workshop looked at bridging the gaps between short-term and long-term perspectives, theoretical approaches and pragmatic solutions, operational challenges and policy issues, and industry and academia. By doing so, we will develop the insight needed to build, evaluate, deploy, operate, and maintain AI-based systems that are truly safe.

The workshop received 33 submissions and accepted 12 papers, an overall acceptance rate of 39 percent. In addition, we invited 5 talks and accepted 10 submissions as short papers for a poster presentation. The workshop program was organized into five thematic sessions. The thematic sessions followed a highly interactive format. They were structured into short pitches and a panel slot to discuss both individual paper contributions and common issues.

The first session discussed safe planning and operation of autonomous systems. Novel approaches were presented to deal with the consequences of planning with reduced-order models and for robust motion planning and safety benchmarking in human work spaces. Finally, a virtualization approach to safe interruptibility, also known as the big red button problem, was presented.

The second session explored new paradigms in AI and artificial general intelligence safety, including a proposal on how to achieve robust end-to-end alignment of AI systems, and the idea of using integrative biologic simulation and neuropsychology in AI safety problems.

The third session focused on safety for automated driving, exploring the operational design domains, objects, and events that must be considered to ensure safety, as well as safety monitoring techniques with crash prediction networks.

The fourth session covered the specification of safety-relevant requirements in AI and machine learning and surveyed safety-relevant characteristics both in near-term and long-term AI safety problems.

The final session discussed adversarial machine learning and featured outstanding contributions on detecting backdoor attacks by activation clustering, object detectors for adversarial patch attack, and accountability assurance by using adversarial machine learning.

The workshop also featured inspirational speakers. Two keynotes opened the morning and afternoon sessions. Sandeep Neema (DARPA) talked about DARPA's assured autonomy program, and Francesca Rossi (IBM) spoke about ethically bounded AI. Three additional invited talks discussed important challenges in AI safety. Peter Eckersley (Partnership on AI) discussed the impossibility and uncertainty theorems in AI value alignment. Ian Goodfellow (Google Brain) presented his insightful views on adversarial robustness for AI safety. Finally, Alessio R. Lomuscio (Imperial College London) presented techniques for reachability analysis in neural agent–environment systems.

The workshop was cochaired by Huáscar Espinoza, José Hernández-Orallo, Seán Ó hÉigeartaigh, Xiaowei Huang, and Mauricio Castillo-Effen, all of whom also submitted reports. The papers were published as *CEUR Workshop Proceedings*, Volume 2301.

Dialog System Technology Challenge

The Dialog System Technology Challenge has been a premier research competition for dialogue systems since its inception in 2013. This AAAI-19 workshop is the seventh in a series of challenges focusing on end-to-end dialogue tasks, to explore the issue of applying end-to-end technologies to dialogue systems in a pragmatic way.

The challenge consists of three tracks: noetic end-to-end response selection, end-to-end conversation modeling, and audiovisual scene-aware dialogue.

The Noetic End-to-End Response Selection challenge consists of subtasks on two data sets, one focused but small (course advising) and the other more diverse but large (Ubuntu support). In each, participants select the correct next utterances from a set of candidates and even indicate that none of the proposed utterances is a good candidate. The objective is to push utterance classification toward real-world problems.

The End-to-End Conversation Modeling: Moving Beyond Chitchat – Sentence Generation track proposes an end-to-end conversational modeling task, where the goal is to generate conversational responses that go beyond chitchat, by injecting informational responses that are grounded in external knowledge.

The Audio Visual Scene-Aware Dialogue track proposes an end-to-end audiovisual scene-aware dialogue system, where the goal is to understand scenes to have conversations with users about the objects and events around them.

More than 220 participants were registered, and about 40 teams participated in the final challenge. We had a one-day wrap-up workshop at AAAI-19 to review the state-of-the-art systems, share novel approaches to the challenge tasks, and discuss future directions for dialogue technology. We had about 80 preregistrations for the workshop, and more participants joined onsite. We accepted 32 papers reporting the systems submitted to the Dialog System Technology Challenge and accepted 3 general technical papers for dialogue technologies. Holger Schwenk from Facebook was invited as a keynote speaker and talked about massively multilingual dialogue and Q&A. To initiate the next challenge, we had a session to introduce the seven track proposals. The poster session, which included lunch, had three sponsors.

Chiori Hori and Koichiro Yoshino organized the 2019 challenge, and Seokhwan Kim is a member of the steering committee. All three submitted this report.

Engineering Dependable and Secure Machine Learning Systems

Contemporary software systems increasingly encompass machine learning components. Similarly to other software systems, machine-learning-based systems must meet dependability, security, and quality requirements.

Standard notions of software quality and reliability such as deterministic functional correctness, black-box testing, code coverage, and traditional software debugging may become irrelevant for machine-learning systems, because of their nondeterministic nature, reuse of high-quality implementations of machine-learning algorithms, and lack of understanding of the semantics of learned models, for example, when deep learning methods are applied. Thus novel methods are called for, as well as new methodologies and tools to address quality and reliability challenges of machine-learning systems.

Broad deployment of machine-learning software in networked systems inevitably exposes the software to attacks. While classic security vulnerabilities are relevant, machine-learning techniques have additional weaknesses, some already known (for example, sensitivity to training data manipulation) and some yet to be discovered. Hence, there is a need for research as well as practical solutions to machine-learning security problems.

The Engineering Dependable and Secure Machine Learning Systems workshop focused on such topics. It included original contributions exposing problems and offering solutions related to dependability, quality assurance, and security of machine learning systems. It combined several disciplines, including machine learning, software engineering (with emphasis on quality), security, and algorithmic game theory. It also promoted a discourse between academia and industry in a quest for well-founded practical solutions.

The workshop was a well-attended, lively meeting of researchers from academe and industry. The presentations, and the discussions that followed, were very fertile and inspiring, leading to new ideas for future research on adversarial, reliable, and secure machine learning.

The Engineering Dependable and Secure Machine Learning Systems workshop was organized by Eitan Farchi (IBM Research), Onn Shehory (Bar Ilan University), and Guy Barash (Western Digital). Links to workshop papers are available on the workshop's website. Revised versions will be published in a special issue of the journal *Software Testing, Verification and Reliability* on adversarial machine learning. This report was submitted by Eitan Farchi, Onn Shehory, and Guy Barash.

Games and Simulations for Artificial Intelligence

Games have a long history in AI research, dating back to at least 1949, when Claude Shannon (shortly after developing information entropy) got interested in writing a computer program to play the game of chess. Since then, there has been enduring interest in creating computer programs that can play games as skillfully as human players, even beating respective world champions. The progress over the last 70 years has included two-player board games such as checkers, backgammon, chess and Go; 2D Atari games; and 3D video games such as Doom, Starcraft II, and Dota 2.

It's not just games that have played a central role in AI development. Game engines themselves (and other simulation platforms) are becoming a powerful tool for researchers across many disciplines. A large number of platforms have recently been created to study such research problems as playing video games, physics-based control, locomotion, 3D pose estimation, visual navigation, natural language instruction following, embodied question answering, and autonomous vehicles (for example, Arcade Learning Environment, General Video Game AI, Allen Institute AI2-Thor, Facebook Habitat, Microsoft AirSim, and Unity ML-Agents Toolkit).

A primary reason for adopting game engines in AI research is their ability to generate large amounts of synthetic data. This ability is especially profound in scenarios in which data-set generation in the real world is prohibitively expensive or dangerous. Another reason is their rendering quality and physics fidelity, which enable the study of real-world problems in a safe and controlled environment. It also enables models trained on synthetic data to be transferred to the real world with minimal changes.

We foresee game engines and simulation platforms playing a very important role in AI's future development. As such, we were interested in hosting a workshop that brings researchers across AI who are either using simulations platforms or interested in learning more. The full-day workshop was organized into three sessions: Games and Environments, Autonomous Vehicles and Robotics, and Vision and Language. Each session had two invited speakers. The workshop also reviewed 20 paper submissions, 7 of which were accepted into the workshop.

The invited speakers presented a wide range of topics anchored in how simulation platforms are used to better train and evaluate AI systems. This included an overview of games and simulation platforms (Danny Lange); a discussion of an extensive, lightweight, and flexible framework, ELF, for game research at Facebook AI (Yuandong Tian); an overview of how video games have evolved as a benchmark for learning and planning, plus a discussion on the next 5 years (Julian Togelius); a real-world case of using Microsoft AirSim to develop autonomous agents for the open world, plus an example of sim-to-real learning (Shital Shah); a deep dive into OpenAI's Dactyl system and its use of high-performance rendering as a back end for robotics applications (Maciek Chociej); an overview of the core components needed to build an embodied 3D simulation platform, which lead to the development of Facebook AI Habitat (Manolis Savva); and finally, a discussion on the study of natural language understanding within simulated environments (Yoav Artzi).

Marwan Mattar, Roozbeh Mottaghi, Julian Togelius, and Danny Lange served as cochairs of the workshop and submitted this report. The accepted papers were published as arXiv preprint1903.02172.

Health Intelligence

Population health intelligence includes a set of activities to extract, capture, and analyze multidimensional socioeconomic, behavioral, environmental and health data to support decision making to improve the health of various populations. Advances in AI tools and techniques and Internet technologies are dramatically changing how scientists collect data and how people interact with each other and with their environment. The Internet is also increasingly used to collect, analyze, and monitor health-related reports and activities and to facilitate health-promotion programs and preventive interventions. In addition, to tackle and overcome several issues in personalized healthcare, information technology will need to evolve to improve communication, collaboration, and teamwork among patients, their families, healthcare communities, and care teams involving practitioners from different fields and specialties.

This workshop follows the success of previous health-related AAAI workshops, including those focused on personalized and population healthcare, and the two subsequent joint workshops held at AAAI-17 and AAAI-18. This year's two-day workshop brought together a wide range of participants (roughly 70 registrants) from the multidisciplinary field of medical and health informatics. Participants were interested in the theory and practice of computational models of web-based public health intelligence as well as personalized healthcare delivery. The full and short papers and the posters presented at the workshop covered a broad range of disciplines within AI, including knowledge representation, machine learning, natural language processing, prediction, mobile technology, inference, and dialogue systems. From an application perspective, presentations addressed topics in epidemiology, environmental and public health informatics, disease surveillance and diagnosis, medication dosing, health behavior monitoring, and human-computer interaction.

The workshop included an invited talk by Barry O'Sullivan (University College Cork), who gave a presentation on case studies in improving healthcare delivery. To further promote the work presented at the workshop, the authors of mature research were given the opportunity to submit revised and significantly extended manuscripts for review to appear in a special issue of the *Journal of Artificial Intelligence in Medicine* on precision digital medicine and health. Martin Michalowski, Arash Shaban-Nejad, David L. Buckeridge (McGill University), John S. Brownstein (Harvard University), and Niels Peek (University of Manchester) will serve as guest editors for this collection.

Martin Michalowski and Arash Shaban-Nejad served as cochairs of this workshop and submitted this report. The workshop papers were published by Springer in the *Studies in Computational Intelligence* series.

Knowledge Extraction from Games

The second workshop on Knowledge Extraction from Games again focused on mechanically extracting knowledge from games — including but not limited to game rules, character graphics or audio, environments, high-level goals or heuristic strategies, transferrable skills, aesthetic standards and conventions, or abstracted models of games. Games enjoyed by human players have been an area of interest for AI from the days of fraudulent chess automata to today's superhuman play of Go and Atari games. But games are more than just planning problems: whereas deep Q-learning and other efforts yield successful policies for playing specific games, we might want to ask different questions of a game system besides how does one win.

Games provide useful structuring information for many reasoning tasks and are therefore an ideal environment for this work. For example, games in which nonplayer characters (or environment design) offer hints to solving problems might be useful stepping stones toward contextual query answering; it is not enough to find the right solution, but one must identify the relationship between the textual or visual hints and the correct embodied actions. Games often share genre conventions and other similarities, or continually force a player to learn new skills or exercise their existing competencies in novel contexts; therefore, it seems especially interesting to explore transfer learning and analogical reasoning within and between games.

This workshop brought together practitioners from these communities and others whose goals overlap but whose approaches are developed in parallel — search, general (video) game playing, reinforcement learning, design support, human factors, sequence analysis, and others. We had a significant increase in submissions, accepted papers, and attendance from the previous year. Our authors presented 11 papers, ranging from applications of local search for game exploration to an analysis of how humans learn game rules. We also hosted two invited talks. Alexander Zook explained some ways in which Blizzard Entertainment's data science team analyzes players' behavioral data to achieve design goals, and Stanford University's Srijan Kumar explored the automated analysis of player gaze in games of deception.

The key strength of this workshop was, as hoped, the integration of multiple communities of AI and automated reasoning researchers and game designers. The questions posed after talks were stimulating, and at least one of the papers has been cited by upcoming work in automated game state-space exploration. We look forward to hosting the workshop again and seeing the new syntheses that emerge in the next round of submissions!

The cochairs of the workshop were Joseph Osborn (Pomona College), Matthew Guzdial (Georgia Tech), and Samuel Snodgrass (Drexel University). The proceedings were published as *CEUR Workshop Proceedings*, Volume 2313. Joseph C. Osborn wrote this report.

Network Interpretability for Deep Learning

The AAAI-19 workshop on Network Interpretability for Deep Learning brought together scientists, engineers, and students in both academic and industrial communities who are interested in opening the black box of deep neural networks and pursuing interpretable knowledge representations. The main theme of the workshop discussion is to build up consensus in the emerging field of interpretable artificial intelligence and, in particular, to clarify the motivation, typical methodologies, prospective trends, and potential industrial values of studying interpretability of deep neural networks.

This workshop included five invited talks. Su-In Lee (University of Washington) introduced the use of interpretable machine learning in medical applications. Tianfu Wu (North Carolina State University) introduced deep compositional grammar networks. Zhanxing Zhu (Peking University) discussed the role of adversarial learning for learning explainable feature representations. Quanshi Zhang (Shanghai Jiao Tong University) discussed core challenges and solutions to feature interpretability and structure interpretability of deep neural networks. Xianglei Xing (Harbin Engineering University) introduced generator networks with and-or grammars.

The workshop received 30 submissions, including both full papers and extended abstracts; 26 papers were delivered at this workshop — 6 oral presentations and 20 poster presentations. In an oral presentation, David Bau introduced a method to quantify the feature interpretability of generative networks and interpretability-based interactive image generation. In another, Fan Bao introduced a method that used adversarial learning to boost the interpretability of neural networks. Jiaoyan Chen presented how to use knowledge graph to explain logic encoded inside deep neural networks. Kyoung-Woon On visualized semantic structures of videos that were encoded in deep neural networks. Kam Who Ng introduced a universal deep logic convolutional network.

In this workshop, participants shared ideas of prospective trends in interpretable deep learning and agreed that they would like to attend future workshops on similar topics.

Quanshi Zhang, Lixin Fan, and Bolei Zhou served as cochairs of this workshop. The papers of the workshop were published on arXiv. This report was written by Quanshi Zhang.

Plan, Activity, and Intent Recognition

The 2019 Plan, Activity, and Intent Recognition workshop was a successful, fruitful, and well-attended event. Plan recognition, activity recognition, and intent recognition all involve making inferences about other actors from observations of their behavior, that is, their interaction with the environment and with each other. The observed actors may be software agents, robots, or humans. This synergistic area of research becomes especially important, as the AI community dives deeper into the challenges that arise in complex multiagent settings and settings where the effective interaction between agents relies on their ability to perform explainable behavior, a behavior that is easily interpreted by an observer. Plan, activity, and intent recognition confronts these challenges by combining and unifying techniques from user modeling, machine vision, intelligent user interfaces, human-computer interaction, autonomous and multiagent systems, natural language understanding, automated planning, and machine learning.

This year's workshop was preceded by a tutorial, held the previous day. The tutorial allowed a lively discussion between participants and attendees that was profound and effective. This was especially important given the wide range of approaches and application that were presented, including work on all three main threads of research: activity, plan, and goal recognition. The rising interest in plan recognition resulted in the highest submission rate in the more than 10 years of existence of the workshop. Exceptionally high-quality papers were presented, and we would like to thank our workshop organizing committee for their reviewing efforts.

The workshop also had two well-attended talks by eminent AI researchers.

Shlomo Zilberstein (University of Massachusetts Amherst) talked about how plan recognition is an essential component of multiagent decision making. He examined plan recognition in this context and showed that existing algorithms for multiagent planning can perform plan recognition implicitly.

David Smith noted that there has been increasing interest in the generation of behavior that is understandable or interpretable by an observer. In the robotics and planning communities, various notions have been introduced and investigated, including explicability, legibility, predictability, transparency, privacy, security, and obfuscation. Not surprisingly, he noted, many of these notions are related to goal and plan recognition. However, it is not always clear exactly how these notions relate to each other, or what assumptions are being made about the domain model and computational capabilities of the agent and observer. In his talk, he presented a formal taxonomy of different forms of interpretability and uninterpretability and pointed out some interesting variations and combinations that have not yet been considered or explored.

Sarah Keren, Reuth Mirsky, and Christopher Geib cochaired this workshop and submitted this report.

Reasoning and Learning for Human-Machine Dialogues

Although natural conversation has been a key subarea of AI for decades, renewed interest has been fueled by availability of devices suitable for human-technology interaction (like Amazon Alexa, Google Home, and social robots) and new easy-to-use commercial tools (like IBM Watson, Google DialogFlow, and the Microsoft bot framework).

However, beyond basic demonstration, there is little experience in how conversation agents, or chatbots, can be designed and used for real-world applications that need decision making under uncertainty and constraints (for example, sequential decision making) and are invaluable to users.

In techniques, statistical and machine learning methods are well entrenched for language understanding and entity detection, but the wider problem of dialogue management is unaddressed with mainstream tools supporting rudimentary rule-based processing. There is an urgent need to highlight the crucial role of reasoning methods such as constraints satisfaction, planning, and scheduling, and learning. By working together, such methods can build an end-to-end conversation system that evolves over time. From the practical side, conversation systems for working with people are needed that allows the systems to explain their reasoning, convince humans to make choices among alternatives, and meet ethical standards demanded in real-life settings.

To discuss these, the second Reasoning and Learning for Human-Machine Dialogues workshop built on the success of the first event at AAAI 2018 and was attended by more than 60 AI researchers from around the world. The program included four invited talks, six presentations of reviewed full papers, five lightening talks accompanied by posters, a competition on building agents with open data, and a panel discussion on a topical subject.

The day started with an invited talk by Phil Cohen of Monash University titled "Toward Collaborative Dialogue," along with an accompanying paper, "Back to the Future for Dialogue Research." Cohen summarized the role reasoning played in supporting expressivity in previous dialogue systems and how they can extend the limits of current learning-based systems. The second talk was given by Koichiro Yoshino of NAIST, who gave a summary of the Dialog System Technology Challenge and accompanying AAAI 2019 workshop, which was running in parallel. This helped the attendees gain a wider perspective. The third talk by Jim Dewan of IBM, "Using Conversation Agents for Customer Support at Scale — the IBM Case Study," described how conversation agents are being designed and deployed at IBM at production scale using knowledge graphs of products, issues, and their resolutions. The fourth talk was on smart chatbots for enhanced health by Amit Sheth of Wright State University. Sheth described how conversation agents are being used in health care with multisensory sensing, knowledge representation, reasoning, and learning for monitoring a patient, appraising, and intervening in treatment of adherence.

Authors of peer-reviewed papers and posters discussed such ideas as design, selection of objective functions, control of sentiments, and support for proactiveness in neural network–based dialogue systems, and their objective functions for dialogue generation and sentiment control; model-based reasoning for dialogue generation; models of laughter and body movement in rich interaction; and how elders perceive commercial dialogue systems. The papers generated a lot of questions and discussions. The organizers also discussed an initiative, launched with the support of *AI Journal*, to promote building of task-oriented agents that people often encounter (like traffic and subways, or health and diseases) and for which data are commonly available in open data portals. Although there were no entries, the organizers have sponsored early researchers to discuss the topic at the workshop and build open-source chatbot implementations as reference. The day ended with an engaging panel moderated by Imed Zitouni and included Radu Marinescu of IBM, Amit Sangroya of TCS, and the invited speakers Amit Sheth and Phil Cohen on the challenges in quickly building high-quality conversation agents that people actually want to use. A few notable points that arose are that chatbots are currently deployed by organizations for cost reasons but they are not the preferred interaction mode for most customers. This may be due to their lack of capability, value perception, or design gaps. Second, both reasoning and learning techniques have a role to play for effective dialogue management, and we need data sets that exemplify the need for both capabilities.

The event thus continued the momentum from the first event and built on it with a mix of theoretical and practical discussions. The attendees expressed satisfaction, and many told the organizers that a follow-up workshop will be worthwhile to build further research momentum around this topic of significant application potential.

Biplav Srivastava, Susanne Biundo, Ullas Nambiar, and Imed Zitouni served as cochairs of the workshop. The papers have been published by the authors on arXiv. This report was submitted by Biplav Srivastava and Imed Zitouni

Reasoning for Complex Question Answering

Question answering (QA) systems have made rapid progress in the last few years — particularly for simple factoid questions and for machine comprehension tasks where a short reference text that contains the answer is given as input. The Reasoning for Complex Question Answering Workshop initiated a conversation about solving much more complex QA tasks that go beyond the simple factoid and machine comprehension settings. Related subgoals of the workshop were to discuss standardizations within the AI community for complex QA problems and to foster collaboration between the AI and computational linguistics communities.

Complex QA refers to a broad set of QA tasks in which the questions, the answers, or the reasoning process required to arrive at the answer are complex. For example, questions may be long and require inference (such as numerical reasoning questions), or questions may be subjective (for example, “Which restaurant should I eat at?”), or the process of arriving at an answer may be complex (for example, the Winograd challenge). The workshop produced an overall consensus that to solve most complex QA tasks, machines require deeper language and world understanding, as well as deep reasoning capabilities. In essence, the workshop emphasized the importance of traditional AI ideas for the field of modern QA.

The keynote talk was given by Eduard Hovy from Carnegie Mellon University, who set the tone with introspection on whether deep learning is really teaching much intelligence to the machine and whether information encoded in word vectors can be equated to giving the machine enough general intelligence about QA. Hovy emphasized that most neural QA models are learning glorified semantic matching patterns and do not possess the reasoning capabilities needed for complex QA. The talk also discussed the characteristics of data sets, and in particular encouraged the creation and use of data sets that require complex reasoning and might be difficult for existing pattern-matching algorithms to solve. Several case studies where complex answering methodologies are needed were described — for example, one where questions need to be converted into an executable program or script, rather than treated as a pattern to match.

Invited talks by Ashish Sabharwal (Allen Institute for Artificial Intelligence), Chitta Baral (Arizona State University), Kenneth Forbus (Northwestern University), and Michael Witbrock (IBM Research) delved into various kinds of reasoning techniques that can aid complex QA. Sabharwal described the value of multihop reasoning — piecing together multiple sources of information to arrive at an answer. Baral focused his talk on commonsense knowledge and reasoning, specifically the role of traditional knowledge representation in such tasks. Forbus introduced analogical reasoning as a promising way to train QA systems in a data-efficient manner. Witbrock’s talk touched on the use of external knowledge, and mathematical reasoning as an important aspect of reasoning for QA.

With respect to community standardization, Witbrock mentioned the Mizar math library as a challenge task. Existing theorems need to be selected and chained to build the requisite proof chain for the theorem at hand. Sabharwal’s talk highlighted the importance of leaderboards, and how they spur research on a given task and help track the overall community progress on it.

At the end, the panel brought back all the invited speakers for a lively conversation about the current state of QA and paths toward progress. Important comparisons were drawn between humans and machines: whereas humans can answer new QA problems even when they have not seen any questions from the new data set, machine-learning systems are brittle and need to be trained on each data set separately. Similarly, although humans can explain their reasoning process for an answer, machines thus far have not been endowed with such capabilities. Overall, there was a sense of dissatisfaction at the current limited success of purely data-driven machine-learning systems, and the conversation called for richer semantic representations, with AI reasoning methods as a path toward substantial progress.

In addition to the invited talks and the panel, the workshop featured presentations on eight peer-reviewed technical papers, which delved into various aspects of complex QA. Overall, the workshop was extremely well attended — registration closed early because of room capacity. The workshop was coorganized by Kartik Talamadupula, Peter Clark, Rajarshi Das, Pavan Kapanipathi, Mausam, and Michael Witbrock. This report was submitted by Mausam, Kartik Talamadupula, Peter Clark, and Pavan Kapanipathi.

Recommender Systems and Natural Language Processing

The interdisciplinary Recommender Systems and Natural Language Processing workshop is at the intersection of recommender systems and natural language processing. The primary goal of this workshop was to identify common ideas and techniques that are being developed and used in both disciplines and to further explore the synergy between the two. While at first glimpse these research fields may seem independent of each other, a new field is emerging where the two meet, especially when dealing with data like consumer reviews; combining ideas from recommender systems and natural language processing allows problems like recommendation, sentiment analysis, or question answering to be solved with higher fidelity, interpretability, and so on.

The two main types of intersection between the two research fields were addressed in this workshop. The first was on the algorithmic side, where both research fields have been inspired by ideas that originated in each other’s field. For example, session-based recommender systems use recurrent neural networks to model complex user-session context data. On the other hand, several recent natural language processing (NLP) techniques, such as word embeddings, have been shown to have strong roots in matrix factorization algorithms, which are the bread and butter of modern recommender systems.

The second intersection type arises in topics that inherently combine both disciplines. Examples of such topics are recommendations of textual items, utilization of user reviews to improve recommendations, conversational (dialogue-based) recommenders, generation of natural language explanations of recommendations, and reading comprehension of reviews to infer relationships between products.

The primary purpose of this workshop was, therefore, to encourage more fundamental interdisciplinary research on recommender systems and NLP. We leveraged this workshop to promote such research by bringing together strong researchers from both communities. Our goal was to encourage researchers to conduct joint work and exchange ideas and knowledge that can be used to solve problems in both fields.

Oren Sar Shalom, Vahid Noroozi, Mengting Wan, and Julian McAuley served as chairs of the workshop and submitted this report.

Reinforcement Learning in Games

Games have been used as benchmarks for rational decision making since the beginning of AI. Although the games and search community developed separately from the reinforcement learning community, the synthesis of techniques from these two research areas has led to groundbreaking results: agents learning to play at an expert level from just the rules of the game and simulated interaction. Recent successes of deep reinforcement learning have sparked new interest in self-play reinforcement learning in games. The goal of this workshop was to bring together the members of these communities to discuss some of the challenges, ideas, and potential next steps.

The workshop started with an hour-long mini-tutorial given by the organizers, outlining some of the historical perspectives, foundations, algorithms, and approaches to the various settings: turn-taking perfect information games, simultaneous (Markov) games, and imperfect information games. The workshop included 39 accepted papers split into 3 poster sessions and 4 oral presentations. Invited talks were given by Georgios Piliouras (Singapore University of Technology and Design) on learning dynamics in games, Emilie Kaufmann (CNRS) on bandit approaches in Monte Carlo tree search, and Michael Littman (Brown University) on learning in general-sum environments.

The workshop closed with a panel discussing some of the key issues, limitations of current techniques, and outstanding problems. We were lucky to collect an outstanding list of panelists that included many principal people of the historical milestones introduced at the very beginning of the workshop.

Major themes of the workshop were convergence analyses, learning dynamics, and how to extend beyond the case of two-player, zero-sum games. The first talk presented results on tabular value iteration for a three-player zero-sum game, showing cyclic learning behavior in some cases. Results on gradient-based learning in continuous games and transfer learning in cooperative games followed, ending with a talk on supervised learning for Skat, an imperfect information card game. These themes were also well reflected by the posters, which also included new benchmarks and competitions, scaling to larger environments, exploration, learning from demonstrations, and approaches to difficult single-agent problems such as Atari, bin packing, and combinatorial optimization.

The workshop attracted more than 100 participants having various areas of expertise. A brief informal poll at the end revealed that the workshop managed to bring together people from both reinforcement learning and game theory backgrounds.

There was a significant level of interaction among participants and some indication that future workshops on this topic would be of interest. Marc Lanctot, Julien Pérolat, and Martin Schmid served as cochairs of the workshop.

Reproducible AI

AI, like any science, must rely on reproducible experiments to validate results. Lately, many researchers have noticed and reported that reproducing results from empirical AI research is not easily accomplished, or even possible at all, despite the experiments being fully conducted on computers. Some of the issues related to reproducibility are caused by poor experiment design and documentation. However, AI research has its own unique reproducibility challenges related to the use of analytical methods that are actively investigated, problems related to nondeterminism in standard benchmark environments, and variance intrinsic to AI methods.

The objective of the workshop was to facilitate sharing of experiences related to reproducibility, a discussion of what could be done to combat the reproducibility problems, and making a roadmap for improving the reproducibility of research results in AI.

Experiences were shared mainly through eight invited talks that presented platforms for simplifying reproducibility, theoretical ideas, practical experience in reproducing experiments, and correcting misconceptions about reproducibility. Discussions were facilitated through a panel discussion and an active working session, in which everyone attending participated. The outputs of the working session were recommendations and a roadmap for how AAAI could implement these.

Matei Zaharia (Stanford University and Databricks) presented Mlflow, a cloud platform for large-scale data analytics and machine learning that supports experiment tracking, comparing experiments, reusable workflows, and more. Matei argued that reproducibility actually matters more for practitioners than scientific researchers and discussed why developing machine-learning systems is harder than traditional software development. Odd Erik Gundersen presented arguments for why a framework for measuring reproducibility is needed and suggested characteristics it should have.

Joel Grus (Allen Institute for AI) shared his not-so-very-secret opinion of not liking Notebooks because they make reproducibility harder. Many well-formed arguments were made, and good examples of what to do instead were shown. Daniel Garijo (University of Southern California) explained the requirements of the scientific paper of the future. The future entails papers that contain not only text but also data, software, experiment setup, and dependencies, while supporting open science and a digital scholarship. Examples of how this can be achieved were given. Hugo Jair Escalante (The ChaLearn Collaboration) presented work on machine-learning challenges and how these can be used to establish benchmarks and fair comparison among methodologies.

After the lunch break, which was far too short, Peter Bull (DrivenData) talked about how we should apply the lessons learned from 50 years of software development to increase reproducibility. Prabhat Nagarajan (University of Texas at Austin) explained how they were able to achieve deterministic implementations on deep reinforcement learning algorithms. Surprisingly, although deterministic on individual computers, because of the parallelism of GPUs, the results will differ between computers and hence are irreproducible. Yuandong Tian rounded off the presentations by explaining the work at Facebook AI Research on reproducing AlphaZero on the ELF platform. The presentation discussed how superhuman performance was achieved, and a thorough ablation analysis of the system was conducted.

After the presentations, a lively panel discussion was moderated by Yolanda Gil. The panel consisted of Pascal Van Hentenryck, Ashok Goel, and Odd Erik Gundersen, but the audience had many questions and comments as well. The panel discussion started with short introductions. Professor Van Hentenryck presented statistics showing that papers with supplemental material were more likely to be accepted at AAAI 2019 than papers without. He also argued for the controversial view that reproducibility could be linked to publication. Ashok Goel highlighted an issue related to reproducibility and one-shot robot learning from demonstration caused by the variation of the human instructors. Also, Ashok talked about *AI Magazine*'s commitment to reproducibility by introducing a new column on the topic. Among other things, the panel discussed how the research community is reluctant to make reproducibility a key concern although it is a key component of science.

Finally, most of the workshop participants joined in on a working session with the goal of making a roadmap for how to increase the reproducibility of results published by AAAI. The participants worked in four different groups and proposed concrete actions for the various actors in the research community on what they could do. After discussing and making a list of actions in the groups, the four groups presented their proposals to the other participants. One hour on overtime, after an energetic final session, the workshop was concluded. The workshop was organized by the cochairs Yolanda Gil, Joelle Pineau, Satinder Singh, and Odd Erik Gundersen. This report was written by Odd Erik Gundersen.

Guy Barash is affiliated with Western Digital.

Mauricio Castillo-Effen is a senior researcher at Lockheed Martin.

Niyati Chhaya is affiliated with Adobe Research.

Peter Clark is a senior research manager at the Allen Institute for Artificial Intelligence.

Huáscar Espinoza is principal researcher at Commissariat à l´Énergie Atomique, France.

Eitan Farchi is affiliated with IBM Research, Haifa.

Christopher Geib is affiliated with SIFT LLC.

Odd Erik Gundersen is the chief AI Officer at TrønderEnergi AS and an adjunct associate professor at Norwegian University of Science and Technology.

Seán Ó hÉigeartaigh is the executive director of the University of Cambridge's Centre for the Study of Existential Risk and program director at the Leverhulme Centre for the Future of Intelligence.

José Hernández-Orallo is a professor at the Universitat Politècnica de València, Spain.

Chiori Hori is a principal research scientist of Mitsubishi Electric Research Laboratories (MERL), USA.

Xiaowei Huang is a lecturer at the Department of Computer Science, University of Liverpool, UK.

Kokil Jaidka is affiliated with Nanyang Tech University.

Pavan Kapanipathi is a research staff member at IBM Research AI.

Sarah Keren is affiliated with Harvard University.

Seokhwan Kim is a research scientist at Adobe Research, USA.

Marc Lanctot is a research scientist at DeepMind Alberta, Edmonton, Canada.

Danny Lange is vice-president of AI and machine learning at Unity Technologies.

Julian McAuley is an assistant professor at the University of California San Diego.

David Martinez is an associate division head at MIT Lincoln Laboratory.

Marwan Mattar is a senior manager of Machine Learning at Unity Technologies.

Mausam is an associate professor of computer science at Indian Institute of Technology Delhi.

Martin Michalowski is affiliated with the University of Minnesota School of Nursing.

Reuth Mirsky is affiliated with the University of Texas.

Roozbeh Mottaghi is a research scientist at the Allen Institute for Artificial Intelligence.

Joseph Osborn is an assistant professor at Pomona College.

Julien Pérolat is a research scientist at DeepMind London, United Kingdom.

Martin Schmid is a research scientist at DeepMind Alberta, Edmonton, Canada.

Arash Shaban-Nejad is affiliated with the University of Tennessee Health Science Center-Oak Ridge National Laboratory.

Onn Shehory is affiliated with Bar Ilan University.

Biplav Srivastava is a distinguished data scientist at IBM's Chief Analytics Office at Armonk, NY.

William Streilein is a member of the principal staff at MIT Lincoln Laboratory.

Kartik Talamadupula is a research staff member at IBM Research AI.

Julian Togelius is an associate professor in the Tandon School of Engineering, New York University.

Koichiro Yoshino is an assistant professor at the Nara Institute of Science and Technology (NAIST), Japan.

Quanshi Zhang is an associate professor at the John Hopcroft Center and the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University.

Imed Zitouni is a principal research manager of the conversation understanding group at Microsoft.