1-1-2019

Reports of the AAAI 2019 Spring Symposium Series

Ioana Baldini
IBM Research

Clark Barrett
Stanford University

Antonio Chella
Università degli Studi di Palermo

Carlos Cinelli
University of California, Los Angeles

David Gamez
Middlesex University

See next page for additional authors

Recommended Citation

Baldini, Ioana; Barrett, Clark; Chella, Antonio; Cinelli, Carlos; Gamez, David; Gilpin, Leilani H.; Hinkelmann, Knut; Holmes, Dylan; Kido, Takashi; Kocaoglu, Murat; Lawless, William F.; Lomuscio, Alessio; Macbeth, Jamie C.; Martin, Andreas; Mittu, Ranjeev; Patterson, Evan; Sofge, Donald; Tadepalli, Prasad; Takadama, Keiki; and Wilson, Shomir, "Reports of the AAAI 2019 Spring Symposium Series" (2019). Computer Science: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/csc_facpubs/162

Authors

Ioana Baldini, Clark Barrett, Antonio Chella, Carlos Cinelli, David Gamez, Leilani H. Gilpin, Knut Hinkelmann, Dylan Holmes, Takashi Kido, Murat Kocaoglu, William F. Lawless, Alessio Lomuscio, Jamie C. Macbeth, Andreas Martin, Ranjeev Mittu, Evan Patterson, Donald Sofge, Prasad Tadepalli, Keiki Takadama, and Shomir Wilson

Reports of the AAAI 2019 Spring Symposium Series

Ioana Baldini, Clark Barrett, Antonio Chella, Carlos Cinelli, David Gamez, Leilani H. Gilpin, Knut Hinkelmann, Dylan Holmes, Takashi Kido, Murat Kocaoglu, William F. Lawless, Alessio Lomuscio, Jamie C. Macbeth, Andreas Martin, Ranjeev Mittu, Evan Patterson, Donald Sofge, Prasad Tadepalli, Keiki Takadama, Shomir Wilson

■ *The AAAI 2019 Spring Symposium Series was held Monday through Wednesday, March 25–27, 2019, on the campus of Stanford University, adjacent to Palo Alto, California. The titles of the nine symposia were Artificial Intelligence, Autonomous Machines, and Human Awareness: User Interventions, Intuition and Mutually Constructed Context; Beyond Curve Fitting — Causation, Counterfactuals and Imagination-Based AI; Combining Machine Learning with Knowledge Engineering; Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness; Privacy-Enhancing Artificial Intelligence and Language Technologies; Story-Enabled Intelligence; Toward Artificial Intelligence for Collaborative Open Science; Toward Conscious AI Systems; and Verification of Neural Networks.*

AI, Autonomous Machines, and Human Awareness

Applications of machine learning combined with AI algorithms have propelled unprecedented economic disruptions across diverse fields in industry, military, medicine, finance, and others. With the forecast for even larger impacts, the present economic impact of machine learning is estimated in the trillions of dollars. But as autonomous machines become ubiquitous, recent problems have surfaced. Early on, and again in 2018, Judea Pearl warned AI scientists they must "build machines that make sense of what goes on in their environment," a warning still unheeded that may impede future development. For example, self-driving vehicles often rely on sparse data; self-driving cars have already been involved in fatalities, including a pedestrian; and yet machine learning is unable to explain the contexts within which it operates.

ISSN 0738-4602

We propose that these seemingly unrelated problems require an interdisciplinary approach to address Pearl's warning. At our symposium, for example, papers were presented by AI computer scientists, engineers, social scientists, lawyers, physicians, entrepreneurs, philosophers, and others, who addressed how user interventions may explain the mutual context for autonomous machines operating in unfamiliar environments or when experiencing unanticipated events; how autonomous machines can be taught to explain shared contexts by reasoning, inferences or causality and decisions by humans relying on intuition; and how human-machine teams may interdependently affect human awareness, other teams, and society and how these teams may be affected in turn. In short, can context can be mutually constructed and shared between machines and humans to enhance performance? For example, in the Uber accident that killed a pedestrian in 2018, the car, which detected the pedestrian 5 seconds before the human driver did, was a poor team player that did not alert its human-operator teammate when it easily could have.

By extension, we remain interested in whether shared context follows when machines begin to develop subjective states, somewhat like humans', that allow both to monitor and report on their joint interpretations of reality, forcing scientists to rethink the general model of human social behavior and thus elevating the value of an interdisciplinary approach. If dependence on AI and machine learning continues to grow, we and the public are also interested in what happens to context shared by human-machine teams or society when these teams malfunction. As we think through this change in human terms, our ultimate goal is for AI to advance the performance of human-machine teams for the betterment of society wherever these teams interact with other human or machine outsiders.

After completing most of our invited and regular presentations over the first two days, on the third day of our symposium, we had an extended joint session with the Privacy-Enhancing Artificial Intelligence and Language Technologies symposium. In this joint session, we discussed how to apply privacy to teams — for example, to the extent possible, the sharing of context among teammates must remain private to enhance trust and the further sharing of private information within a team. That is, what teammates share should not be disclosed outside of the team context unless rules or laws have been violated.

William F. Lawless, Ranjeev Mittu, and Donald Sofge served as cochairs of this symposium and wrote this report.

Beyond Curve Fitting: Causation, Counterfactuals, and Imagination-Based AI

AI and machine learning have received increased attention from the general public, primarily because of the successful application of deep neural networks to many tasks, including computer vision, natural language processing, and game playing. However, despite all this progress, there is a growing segment of the scientific community that questions whether these successes can be extrapolated to create general AI without a major retooling. The goal of this symposium was to bring together researchers across multiple disciplines (computer science, cognitive science, economics, medicine, statistics) to discuss the capabilities of current AI and machine-learning technologies and the integration of causal and counterfactual reasoning into the data-driven sciences to help alleviate their shortcomings.

The symposium featured a keynote talk by Judea Pearl, professor of computer science and statistics at the University of California, Los Angeles, and Turing Award winner for his work on probabilistic and causal reasoning. Pearl's talk focused on the foundations and types of causal inference in terms of a three-layer causal hierarchy that sharply distinguishes (1) associations, (2) interventions, and (3) counterfactuals. These theoretical foundations unveil how several important problems found throughout society are beyond the reach of the current generation of machine-learning systems but that can be solved with the tools of causal inference.

The symposium was organized into sessions following the format "causality + *x*," where *x* is a select area that included (1) computer vision and imagination, (2) machine learning and AI, (3) the social sciences and economics, and (4) the health sciences. The speakers were asked to discuss the present and future of their fields, including the recent advances in causal inference that changed their areas (in terms of both methodology and practice) as well as the most pressing issues that causal inference tools may be able to help with in the next few years.

Specifically, the session on computer vision and imagination included topics ranging from the construction of causal variables using unsupervised learning to leveraging causal models for enabling unsupervised learning techniques (such as GANs) to sample from interventional distributions. The speakers for the session were Frederick Eberhardt (Caltech), Murat Kocaoglu (MIT-IBM Watson AI Lab), and Mohammed Elhoseiny (KAUST).

The machine learning and AI session featured a discussion on how causal and counterfactual reasoning guides human cognition and decision making, its relationship with reinforcement learning, and initial explorations of how it can be related to deep learning. The speakers for the session were Tobias Gerstenberg (Stanford University), Thomas Dietterich (Oregon State University), and Yoshua Bengio (University of Montreal).

The social sciences and economics session focused on how causal modeling tools can help tackle well-known problems such as confounding bias, selection bias, generalizability of experimental findings, and transportability, as well as several methodological issues that still need to be addressed, mainly involving new challenges related to social interactions and the availability of unstructured and high-dimensional data. The speakers were Kosuke Imai (Harvard University) and Paul Hünermund (Maastricht University).

Finally, talks on the health sciences revolved around how causal modeling has helped clarify long-standing issues in epidemiology, as well as the risks of bias and the (un)fairness of predictive algorithms. The speakers were Maria Glymour (UCSF) and Mark Cullen (Stanford University).

The symposium also held sessions of short talks and posters for the contributed papers over a broad variety of topics, such as bias analysis, causal discovery, missing data, instrumental variables, transportability, counterfactual reasoning, and data fusion. The participants discussed how the recent growth in popularity of machine-learning techniques in their fields has rekindled interest in understanding their theoretical limitations through the lens of causality, and they exchanged ideas with experts from various fields to put this discussion in a broader context. Participants agreed on several challenges that remain to be addressed through the development of new methodological tools, which should be discussed in future symposia.

Elias Bareinboim, Prasad Tadepalli, Sridhar Mahadevan, Csaba Szepesvari, Bernhard Scholkopf, and Judea Pearl served as cochairs of this symposium. Carlos Cinelli, Murat Kocaoglu, and Prasad Tadepalli wrote this report.

Combining Machine Learning with Knowledge Engineering

The AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering aimed to combine machine learning with knowledge engineering. Machine learning helps to solve complex tasks based on real-world data instead of pure intuition. It is most suitable for building AI systems when knowledge is not known or when knowledge is tacit.

Many business cases and real-life scenarios using machine-learning methods, however, demand explanations of results and behavior, particularly when decisions can have serious consequences. Furthermore, application areas such as banking, insurance, and medicine are highly regulated and require compliance with laws and regulations. This specific application knowledge cannot be learned but needs to be represented, which is the area of knowledge engineering.

Knowledge engineering, on the other hand, is appropriate for representing expert knowledge, which people are aware of and which has to be considered for compliance reasons or explanations.

Knowledge-based systems that make knowledge explicit are often based on logic and thus can explain their conclusions. These systems typically require a higher initial effort during development than systems that use machine-learning approaches. However, symbolic machine- learning and ontology-learning approaches show promise for reducing the effort of knowledge engineering.

Because of their complementary strengths and weaknesses, there is an increasing demand for the integration of knowledge engineering and machine learning. Conclusively, recent results indicate that explicitly represented application knowledge could help data-driven machine-learning approaches to converge faster on sparse data and to be more robust against noise.

More than 70 participants of the Combining Machine Learning with Knowledge Engineering AAAI symposium contributed to intense discussion during presentation of 28 position papers and full papers and four poster sessions and demonstrations. Topics covered such application domains as health care, drug development, social networks, material sciences, fake news detection, and product recommendations. The presentations typically focused primarily on either machine learning or knowledge-based systems. However, there was a strong commitment to the importance of combining machine learning with knowledge bases. Focusing on only one aspect will not exploit the full potential of AI.

The participants had the opportunity to attend several keynotes. On the first day, Doug Lenat emphasized a need for a more expressive logic language in his keynote presentation. He gave a recap on the Cyc knowledge base and showed ways to connect knowledge-based systems with machine learning. On the second day, Frank van Harmelen showed the limitations of machine learning, in particular in areas where not much knowledge is available, like the recognition of rare diseases. He introduced the concept of boxology to represent the reusable architectural patterns for combining learning and reasoning. In the plenary session on day two, Aurona Gerber gave a short and witty overview of the AAAI-MAKE symposium by using an analogy to Asterix. On the final day, cochairs Knut Hinkelmann and Andreas Martin concluded the symposium and emphasized that this new joint community should continue contributing on the topic of combining their fields. There was consensus that the topic is worth exploring in the future.

Andreas Martin, Knut Hinkelmann, Aurona Gerber, Doug Lenat, Frank van Harmelen, and Peter Clark were part of the organizing team of this symposium and served as session chairs. The papers of the symposium were published as *CEUR Workshop Proceedings*, Volume 2350. This report was written by Andreas Martin and Knut Hinkelmann.

Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness

The AAAI 2019 spring symposium on Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness discussed interpretable AI for well-being. Interpretable AI is a method and system from which outputs can be easily understood by humans. Especially in the human health and wellness domains, wrong predictions could affect critical judgments in life-or-death situations. AI-based systems must be well understood.

One of the important issues in understanding machine intelligence in 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 social media platforms and commercial recommendation systems being used to manipulate people's inherent cognitive bias.

Another important issue is social embeddedness. AI systems will be deeply embedded in society, and we need to understand how AI is perceived at the society level. Social embeddedness topics include the role of AI in future economics (basic income, impact of AI on GDP) and the well-being of society (happiness of citizens, life quality).

Our symposium included four invited talks to provide new perspectives on interpretable AI for well-being. 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. Judea Pearl (UCLA) introduced the foundations of causal inference. Sidharth Goel (Google AI) introduced DeepVariant, deep learning for genomic variant calling. The final speaker was Peter Pirolli (Florida Institute for Human and Machine Cognition), who gave a talk on interpretable AI for well-being using mobile health in the context of cognitive science.

The symposium technical presentations included 25 papers and 3 posters and demonstrations. Presentation topics included explainable AI, interpretable AI, social embeddedness, cognitive bias, and well-being AI. Takashi Kido (Preferred Networks) presented on limitations of current technologies based on machine learning and discussed the challenges for interpretable AI for well-being. Amy Ding (Carnegie Mellon University) proposed a model of unbiased and explainable algorithmic decision making that treats everyone fairly. Umang Bhatt (Carnegie Mellon University) proposed the idea of temporal explanations as a medical narrative. Sadeq Rahimi (Harvard University) discussed extended mind, embedded AI, and the barrier of meaning. Morteza Shahrezay and Orestis Papakyriakopoulos (Bavarian School of Public Policy at the Technical University of Munich) reported research on estimating the political orientation of Twitter users. Ziehui Leng (University of Tokyo) presented a cross-lingual analysis on culinary perceptions to understand cross-cultural differences. Yuichi Yoda (Ritsumeikan University) reported on a study of basis of AI-based information systems in the case of the AI shogi system Ponanza.

Takashi Kido and Keiki Takadama served as cochairs of this symposium and wrote this report. The symposium papers will be published online as a CEUR workshop proceedings.

Privacy-Enhancing AI and Language Technologies

Privacy remains an evolving and nuanced concern of computer users, as new technologies that use the web, smartphones, and the Internet of Things collect myriad personal information. Rather than viewing AI and human language technologies as problems for privacy, the goal of this symposium was to flip the script and explore how AI and human language technology can help meet a user's desire for privacy when interacting with computers. This event was a successor to Privacy and Language Technologies, a previous AAAI Symposium held in fall 2016.

We focused on two flexibly defined research questions: How can AI and human language technologies preserve or protect privacy in challenging situations? and, How can AI and human language technologies help interested parties (for example, computer users, companies, regulatory agencies) understand privacy in the status quo and what people want?

Talks by the keynote speakers followed these two themes. Jessica Staddon (Google) spoke on opportunities for AI and human language technologies in security and privacy incident management and discovery, leading to a discussion on opportunities for AI and human language technologies to improve how companies and large institutions manage breaches in privacy and security. Serge Egelman (ICSI) gave a talk on empowering users to make privacy decisions in mobile environments, which led to a discussion of how AI and human language technologies can better connect smartphone users with information on how their personal data are shared and collected.

The symposium program also consisted of oral presentations of accepted papers, discussion forums, and a poster session. Privacy policies of apps and websites were a major theme: several participants presented work on improving the usability of privacy policies by extracting key information from them automatically. Other presenters addressed privacy in online social networks and privacy-preserving machine learning. Additionally, the symposium included a joint session with the AI, Autonomous Machines, and Human Awareness symposium to explore potential collaborations.

The symposium lead organizer was Shomir Wilson (Pennsylvania State University), and the coorganizers were Sepideh Ghanavati (University of Maine), Kambiz Ghazinour (Kent State University), and Norman Sadeh (Carnegie Mellon University). The papers of the symposium were published as a CEUR workshop proceedings. This report was written by Shomir Wilson.

Story-Enabled Intelligence

The ability to tell and understand stories draws on many aspects of intelligence, from language understanding, to planning, to commonsense reasoning, mental modeling, creativity, and moral evaluation. As such, aspects of narrative intelligence turn up in various subfields of artificial intelligence and have even sparked subfields of their own. In convening this symposium, our aim was to provide a forum to bring these aspects together. Our watchword was story-enabled intelligence, the principle that the mechanisms that enable humans to understand and tell stories fundamentally enable many other aspects of intelligent behavior.

Accordingly, the symposium gathered researchers from overlapping fields such as planning, narratology, cognitive science, robotics, user interface design, machine learning, and linguistics. The papers and invited talks necessarily covered a broad range of topics but tended to center on one of two key themes.

The first theme that emerged was the explanatory role of story intelligence: story intelligence as a tool for rendering, for example, machine judgments, automated plans, or user interfaces intelligible to human users. For example, our panel on public perception of AI discussed how narratives can foster trust in otherwise opaque systems and what science journalists can do to inform the public on issues relating to AI. Kelly Neville (Soar, Inc.) demonstrated a narrative-based support tool for augmenting the judgments of human analysts; Cindy Bishop (MIT Media Lab) discussed how art can be used to depict and communicate about AI systems; Ted Selker showed how to improve human-computer interfaces using systems that recognize users' storied actions; Joshua Grossman (Stanford) demonstrated conversational tutoring agents in mathematics; and Ron Petrick (Heriot-Watt University) argued that effective, real-time planning systems must sense and adapt to the emotional responses of its users.

The second theme that emerged was how modeling story-understanding mechanisms can shed light on various forms of human intelligence, such as plan understanding, narrative comprehension, and social reasoning. Pat Langley (Institute for the Study of Learning and Expertise) discussed real-time planning in the context of disaster response; Danielle Olson (Massachusetts Institute of Technology) described how to tailor virtual reality narratives to the experiences and biases of individual users; Risto Miikkulainen (University of Texas at Austin) modeled schizophrenic symptoms as breaks in story processing; and Yu-Jung Heo (Seoul National University) argued for richer, gradated data sets for measuring the performance of story understanding systems. Stefan Sarkadi (King's College London), Eugene Shvarts (October), Adam Amos-Binks (Applied Research Associates, Inc.), Mariya Yao (Metamaven), and Jongbin Jung (Stanford) gave talks and organized panel discussions that connected story-enabled intelligence to such wide-ranging topics as argumentation, decision making, trust, deception, and rebellion.

Throughout these discussions, participants weighed the merits of various representations. Rogelio E. Cardona-Rivera (University of Utah), Robert Kirby, Morteza Behrooz (UC Santa Cruz), Andrew Gordon (Institute for Creative Technologies), and Zhutian Yang (Nanyang Technological University) discussed neural networks, word embeddings, scripts, frames, ontological models of narrative, conceptual primitives, and novel forms of causal reasoning and alignment-based learning through stories. John Mitros (University College Dublin), Mary Ellen Foster (University of Glasgow), and Taisuke Akimoto (Kyushu Institute of Technology) focused on interpretability, explainability, and narrative generation.

On reflection, Mark Finlayson (Florida International University) set the overarching agenda in the opening talk, outlining the history and limits of narrative fundamentalism in AI and calling for interdisciplinary research necessary to support such an enterprise. The resulting debate on whether narrative intelligence is fundamental or merely epiphenomenal resonated through the subsequent talks.

Starting from our deliberately all-embracing title, Story-Enabled Intelligence, participants displayed a variety of concrete applications — including planning, visualization, interpretability, computer games, autonomous robots, and art. Topics spanned theory of mind, interpretability as storytelling, prospective cognition, and natural language generation for robotics. By bringing together these diverse applications and fostering a shared vision of narrative-based intelligence, the symposium was able to take the field to another level. Through a variety of talks, conversations, and panel debates, we articulated the fundamental role of stories in promoting better interaction between computer agents and their human users and in developing computational models that help us humans better understand ourselves.

The symposium was organized by Leilani H. Gilpin (Massachusetts Institute of Technology), Dylan Holmes (Massachusetts Institute of Technology), and Jamie C. Macbeth (Smith College). All three prepared this report. Papers from the symposium are being prepared for publication in a *CEUR Workshop Proceedings*.

Toward AI for Collaborative Open Science

The scientific community is undergoing a far-reaching shift toward greater openness and interconnectivity. This trend is driven by a confluence of forces. Spurred by the replication crisis in several branches of science, scientists now place greater emphasis on research transparency at every stage of the scientific process. For example, it is becoming more common to publish preregistered study designs, data sets, data analysis code, preprint articles, and other nontraditional research artifacts. With the emergence of the open access and citizen science movements, scientific research is also becoming more democratic. Finally, data sets are becoming larger and more complex, because of new high-throughput measurement techniques and the possibility of large-scale observational studies. Sharing large, rich data sets is especially important because their full value can rarely be extracted by a single study. New cloud platforms are being developed to support collaborative data science. We expect these trends to continue apace. Future science will be more open, networked, and machine driven than ever before.

The paradigm of collaborative, open science promises to expand human knowledge in new ways, but it also poses serious challenges. The exponential growth of the scientific literature makes it increasingly difficult for researchers to stay current in their own fields, let alone in adjacent fields. The most pressing scientific problems demand collaboration and knowledge sharing between diverse groups of people, from natural and social scientists to engineers and data scientists to policy makers and civil society. Openness in science may be an end in itself, but if its potential value to humankind is to be fully realized, we must find new, systematic ways of making sense of all the newly available scientific artifacts. In this respect, open science remains in its infancy.

We believe that AI has an important role to play in creating, curating, and structuring scientific knowledge and collaboration. It is implausible that AI will supplant human scientists in the near future, but AI could usefully supplement human intelligence, especially in areas where humans perform poorly. For instance, AI agents might help humans discover new, relevant papers within the flood of academic literature or might themselves mine the literature for unknown connections. Such connections are likely to exist, particularly between fields that communicate infrequently. AI could also support human collaboration, perhaps by connecting workers with complementary expertise or by improving the efficiency of knowledge sharing through automation.

Realizing this vision will require sustained collaboration between AI researchers and the broader scientific community. Accordingly, the symposium brought together researchers in basic science, computer science, statistics, and biomedicine, among other areas of academia, industrial research, and the nonprofit sector. The research presented at the symposium ranged from artificial intelligence in its traditional modes, such as knowledge representation and machine learning, to interdisciplinary work on computational tools and collaborative mechanisms for science. Specific themes at the symposium included mining the scientific literature, creating scientific knowledge graphs, data accessibility and reuse, cloud platforms for data science and reproducible research, and crowdsourcing and large-scale collaboration in science.

The diversity in the participants' perspectives and research highlighted the broad, diffuse, and even fragmented status of the open science movement. But it also reaffirmed the importance of venues that bring these different threads together. The problem of integrating AI into the scientific process is inherently interdisciplinary. To solve it, AI researchers must learn from and engage with the larger scientific community.

Ioana Baldini and Evan Patterson served as coorganizers of the symposium and wrote this report.

Toward Conscious AI Systems

Consciousness is part of the physical world, and aspects of it can be studied and potentially replicated by AI systems. Computer models of consciousness can help us to understand biologic consciousness, and the processes at the basis of consciousness may be crudely replicated to build better AI systems. The measurement of consciousness in AI systems is also becoming increasingly important as many people are concerned that AI systems could gain consciousness. Research on conscious AI systems offers outstanding opportunities, but it also brings a set of risks that cannot be underestimated. The AAAI symposium Towards Conscious AI Systems grew out of the AAAI Fall Symposium on AI and Consciousness held in 2007. Since this AAAI meeting, there have been separately organized meetings and workshops on AI and consciousness. However, some of these meetings have been open to limited numbers of invited participants, and there are no regular conferences or workshops. This symposium was an excellent opportunity for people working on conscious AI systems to come together, share their recent research, and reflect on how consciousness relates to AI. The symposium offered a lively space to discuss the connection between AI and fields such as cognitive science, philosophy of mind, ethics, and neuroscience.

Over the past 20 years a wide variety of research projects have been carried out on artificial consciousness, and many types of system have been built. To help us to understand this work, David Gamez (Middlesex University) outlined a classification of the types of machine consciousness, ranging from machines with the same external behavior as conscious systems to machines that are phenomenally conscious. One research theme in the symposium was the capability of AI systems to introspectively analyze their sensory data to extract meaningful reports about their perceptions. A machine may reflect on its perceptions and generate significant reports, as if the machine would experience conscious states. Ron Chrisley (University of Sussex) discussed a sketch of a metacognitive architecture for machine consciousness, and Antonio Chella (University of Palermo) presented a cognitive architecture emulating inner speech in a robot.

Another research theme examined the ways in which studies on the development of consciousness in humans and animals may open new research lines for AI. Henry Shevlin (University of Cambridge) discussed the relationship between general intelligence and consciousness in biologic evolution; David Sahner (EigenMed) proposed a computational testbed for the investigation of evolving agents along the line of developmental robotics, and Minoru Asada (Osaka

University) discussed artificial pain in robots as the basis for empathy, morality, and ethics in a developmental process of consciousness.

Consciousness does not come in isolation, and some talks pointed to the relational aspects of consciousness and their affiliation with theory of mind, attention, and intentionality. Paul Bello (NRL) discussed the relationships between intentional actions and consciousness, and Susmit Jha (SRI International) analyzed the role of shared intentionality as a critical component in conscious AI systems. The tight intermix between logical AI and consciousness dates from a seminal 1995 paper by John McCarthy. Selmer Bringsjord and Naveen Sundar Govindarajulu (Rensselaer Polytechnic Institute) proposed an axiomatic theory of cognitive consciousness that may lead to new reasoning capabilities for machines, along with a new measure of cognitive consciousness.

A theme that spurred lively debated during the symposium was ethical concerns surrounding conscious AI systems operating in society. Of the many talks, both John Sullins (Sonoma State University) and John Murray (San José State University) discussed a system architecture for artificial wisdom in conscious AI systems. The principal argument during the plenary discussion was that the main obstacle to progress in this field is the lack of significant funding, in part because of cultural factors in certain countries and in part because of the lack of cases for possible monetization of the technologies related to conscious AI systems.

Antonio Chella (University of Palermo), David Gamez (Middlesex University), Patrick Lincoln (SRI International), Riccardo Manzotti (IULM University), and Jonathan Pfautz (DARPA) served as cochairs of this symposium. The papers of the symposium were published as *CEUR Workshop Proceedings*, Volume 2287. Antonio Chella and David Gamez wrote this report.

Verification of Neural Networks

The Verification of Neural Networks AAAI symposium was held at Stanford University, Stanford, California, March 25 to 27, 2019. The symposium brought together researchers interested in methods and tools aimed at providing guarantees (formal or otherwise) about the behaviors of neural networks and systems built from them.

Methods based on machine learning are increasingly being deployed for a wide range of problems, including recommender systems, machine vision, and autonomous driving. While machine learning has made significant contributions to such applications, concerns remain about the lack of methods and tools to provide formal guarantees about the behaviors of the resulting systems.

In particular, for data-driven methods to be usable in safety-critical applications, including autonomous systems, robotics, cybersecurity, and cyber-physical systems, it is essential that the behaviors generated by neural networks are well understood and can be predicted at design time. In the case of systems that are learning at run time, it is desirable that any change to the underlying system respect a given safety envelope for the system.

Although the literature on verification of traditionally designed systems is wide and the resulting tools have been successful, there has been a lack of results and effort in this area until recently. The Verification of Neural Networks symposium brought together researchers working on a range of techniques for the verification of neural networks, from formal methods to optimization and testing. One challenge in this upcoming research area is that results are presently being published in several research communities, including formal verification, security and privacy, systems, and AI. The symposium served as a venue for these various communities to interact and build bridges to form a cross-cutting community interested in the verification and validation of systems based on machine learning.

A major theme of the symposium was a focus on specific techniques and tools that have been recently developed for verifying neural networks. Included in this theme were two invited talks by Suman Jana (Columbia University) and Krishnamurthy Dvijotham (DeepMind) as well as two surveys of the field, one by Changliu Liu (Carnegie Mellon University) and one by Nina Narodytska (VMware). Specific tools presented included Marabou, RecurJac, RNSVerify, Sherlock, and Verisig. Mixed integer linear programming was mentioned as a common underlying technique. Other common techniques included abstraction and reachability analysis.

Another theme was robustness. There has been a lot of interest in adversarial robustness of neural networks in the machine learning community generally. Several presentations focused on novel formal methods that can be used to provide guarantees about robustness or to find counterexamples to robustness. It was also stressed how robustness, and derived concepts, provide one basic property that is desirable for many neural networks, particularly in vision applications. This is especially valuable in an area where it is often difficult to obtain specifications for properties that a given neural network should satisfy.

A final major theme was closed-loop systems. Closed-loop systems are multicomponent, often cyber-physical, systems that include a neural network as one component. For such systems, it is important to determine how the neural network fits into and affects the system as a whole. Ideally, verification techniques would guarantee the correct functioning of the entire system, not just the neural network. Various advances in this direction were presented.

The final session of the symposium focused on establishing an effort to collect benchmarks for the community and make them available in a standard format (similar to the SMT-LIB initiative). Armando Tachella (University of Genoa) and Clark Barrett (Stanford) agreed to take a leadership role in this effort.

Clark Barrett and Alessio Lomuscio served as cochairs of this symposium and wrote this report.

Ioana Baldini is a research staff member at IBM Research AI.

Clark Barrett is an associate professor (research) in the Department of Computer Science at Stanford University.

Antonio Chella is a full professor at the Department of Engineering, University of Palermo, Italy.

Carlos Cinelli is a PhD candidate in statistics at the University of California, Los Angeles.

David Gamez is a lecturer at the Department of Computer Science, Middlesex University, London, UK.

Leilani H. Gilpin is a PhD candidate at the Massachusetts Institute of Technology

Knut Hinkelmann is a professor at the FHNW University of Applied Sciences and Arts Northwestern Switzerland.

Dylan Holmes is a PhD candidate at the Massachusetts Institute of Technology.

Takashi Kido is a researcher of Preferred Networks in Japan. He had been a visiting researcher of Stanford University.

Murat Kocaoglu is a researcher at MIT-IBM Watson AI Lab.

William F. Lawless is a professor at Paine College.

Alessio Lomuscio is a professor in the Department of Computing at Imperial College London.

Jamie C. Macbeth is an assistant professor at Smith College.

Andreas Martin is a lecturer and researcher in information systems at the FHNW University of Applied Sciences and Arts Northwestern Switzerland.

Ranjeev Mittu is the Branch Head for the Information Management and Decision Architectures Branch within the Information Technology Division, US Naval Research Laboratory.

Evan Patterson is a PhD candidate in Department of Statistics at Stanford University.

Donald Sofge is a computer scientist and roboticist at the US Naval Research Laboratory.

Prasad Tadepalli is a professor of computer science at Oregon State University.

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

Shomir Wilson is an assistant professor in the College of Information Sciences and Technology at Pennsylvania State University.