Ranjeev Mittu · Donald Sofge
Alan Wagner · W.F. Lawless *Editors*

Robust Intelligence and Trust in Autonomous Systems

Springer

Robust Intelligence and Trust in Autonomous Systems

Ranjeev Mittu • Donald Sofge • Alan Wagner
W.F. Lawless
Editors

Robust Intelligence and Trust in Autonomous Systems

Springer

Editors
Ranjeev Mittu
Naval Research Laboratory
Washington, DC, USA

Donald Sofge
Naval Research Laboratory
Washington, DC, USA

Alan Wagner
Georgia Tech Research Institute
Atlanta, GA, USA

W.F. Lawless
Paine College
Augusta, GA, USA

ISBN 978-1-4899-7666-6 ISBN 978-1-4899-7668-0 (eBook)
DOI 10.1007/978-1-4899-7668-0

Library of Congress Control Number: 2015956336

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

This book is based on the Association for the Advancement of Artificial Intelligence (AAAI) Symposium on "The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems"; the symposium was held at Stanford March 24–26, 2014. The title of this book reflects the theme of the symposium. Our goal for this book is to further address the current state of the art in autonomy at the intersection of RI and trust and to more fully examine the existing research gaps that must be closed to enable the effective integration of autonomous and human systems. This research is particularly necessary for the next generation of systems, which must scale to teams of autonomous platforms to better support their human operators and decision makers.

The book explores the intersection of RI and trust across multiple contexts and among arbitrary combinations of humans, machines, and robots. To help readers better understand the relationships between artificial intelligence (AI) and RI in a way that promotes trust among autonomous systems and human users, this edited volume presents a selection of the underlying theories, computational models, experimental methods, and possible field applications. While other books deal with these topics individually, this book is unique in that it unifies the fields of RI and trust and frames them in the broader context of effective integration for human-autonomous systems.

The volume begins by describing the current state of the art for research in RI and trust presented at Stanford University in the Spring of 2014 (copies of the technical articles are available from AAAI at http://www.aaai.org/Library/Symposia/Spring/ss14-04.php; a link to the presentation materials and photographs of participants is at https://sites.google.com/site/aaairobustintelligence/).

After the introduction, chapter contributors elaborate on key research topics at the heart of effective human-systems integration. These include machine learning, Big Data, workload management, human-computer interfaces, team integration and performance, advanced analytics, behavior modeling, training, and test and evaluation, the latter known as V&V (i.e., verification and validation).

The contributions to this volume are written by world-class leaders from across the field of autonomous systems research, ranging from industry to academia and to

government. Given the diversity of the research in this book, we strove to thoroughly examine the challenges and trends of systems that exhibit RI; the fundamental implications of RI in developing trusted relationships among humans, machines, and robots with present and future autonomous systems; and the effective human systems integration that must result for trust to be sustained.

A brief summary is presented below of the AAAI Symposium in the Spring of 2014.

AAAI-2014 Spring Symposium Organizers

Jennifer Burke, Boeing: jennifer.l.burke2@boeing.com
Alan Wagner, Georgia Tech Research Institute: Alan.Wagner@gtri.gatech.edu
Donald Sofge, Naval Research Laboratory: don.sofge@nrl.navy.mil
William F. Lawless, Paine College: wlawless@paine.edu

AAAI-2014 Spring Symposium: Keynote Speakers

- Suzanne Barber, barber@mail.utexas.edu, AT&T Foundation Endowed Professor in Engineering, Department of Electrical and Computer Engineering, Cockrell School of Engineering, U Texas
- Julie L. Marble, julie.marble@navy.mil, Program Officer: Hybrid human computer systems at Office of Naval Research, Washington, DC
- Ranjeev Mittu, ranjeev.mittu@nrl.navy.mil, Branch Head, Information Management & Decision Architectures Branch, Information Technology Division, US Naval Research Laboratory, Washington, DC
- Hadas Kress-Gazit, hadaskg@cornell.edu, Cornell University; High-Level Verifiable Robotics
- Satyandra K. Gupta, skgupta@umd.edu, Director, Maryland Robotics Center, University of Maryland
- Dave Ferguson, daveferguson@google.com, Google's Self-Driving Car project, San Francisco
- Mo Jamshidi, mo.jamshidi@usta.edu, University of Texas at San Antonio, Lutcher Brown Endowed Chair and Professor, Computer and Electrical Engineering
- Dirk Helbing, dirk.helbing@gess.ethz.ch, http://www.futurict.eu; ETH Zurich

Symposium Program Committee

- Julie L. Marble, julie.Marble@jhuapl.edu, cybersecurity, Johns Hopkins Advanced Physics Lab, MD
- Ranjeev Mittu, ranjeev.mittu@nrl.navy.mil, Branch Head, Information Management & Decision Architectures Branch, Information Technology Division, U.S. Naval Research Laboratory, Washington, DC
- David Atkinson, datkinson@ihmc.us, Senior Research Scientist, Institute of Human-Machine Cognition (IHMC)
- Jeffrey Bradshaw, jbradshaw@ihmc.us; Senior Research Scientist, Institute of Human-Machine Cognition (IHMC)
- Lashon B. Booker, booker@mitre.org, The MITRE Corporation
- Paul Hyden, paul.hyden@nrl.navy.mil, Naval Research Laboratory
- Holly Yanco, holly@cs.uml.edu, University of Massachusetts Lowell
- Fei Gao, feigao@MIT.EDU.MIT
- Robert Hoffman, rhoffman@ihmc.us, Senior Research Scientist, Institute of Human-Machine Cognition (IHMC)
- Florian Jentsch, florian.Jentsch@ucf.edu, Department of Psychology and Institute for Simulation & Training, *Director*, Team Performance Laboratory, University of Central Florida
- Howell, Chuck, howell@mitre.org, Chief Engineer, Intelligence Portfolio, National Security Center, The MITRE Corporation
- Paul Robinette, probinette3@gatech.edu, Graduate Research Assistant, Georgia Institute of Technology
- Munjal Desai, munjaldesai@google.com
- Geert-Jan Kruijff, gj@dfki.de, Senior Researcher/Project Leader, Language Technology Lab, DFKI GmbH, Saarbruecken, Germany

This AAAI symposium sought to address these topics and questions:

- How can robust intelligence be instantiated?
- What is RI for an individual agent? A team? Firm? System?
- What is a robust team?
- What is the association between RI and autonomy?
- What metrics exist for robust intelligence, trust, or autonomy between individuals or groups, and how well do these translate to interactions between humans and autonomous machines?
- What are the connotations of “trust” in various settings and contexts?
- How do concepts of trust between humans collaborating on a task differ from human-human, human-machine, machine-human, and machine-machine trust relationships?
- What metrics for trust currently exist for evaluating machines (possibly including such factors as reliability, repeatability, intent, and susceptibility to catastrophic failure), and how may these metrics be used to moderate behavior in collaborative teams including both humans and autonomous machines?

- How do trust relationships affect the social dynamics of human teams, and are these effects quantifiable?
- What validation procedures could be used to engender trust between a human and an autonomous machine?
- What algorithms or techniques are available to allow machines to develop trust in a human operator or another autonomous machine?
- How valid are the present conceptual models of human networks? Mathematical models? Computational models?
- How valid are the present conceptual models of autonomy in networks? Mathematical models? Computational models?

Papers at the symposium specified the relevance of their topic to AI or proposed a method involving AI to help address their particular issue. Potential topics included (but were not limited to) the following:

Robust Intelligence (RI) topics:

- Computational, mathematical, conceptual models of robust intelligence
- Metrics of robust intelligence
- Is a model of thermodynamics possible for RI (i.e., using physical thermodynamic principles, can intelligent behavior be addressed in reaction to thermodynamic pressure from the environment?)?

Trust topics:

- Computational, mathematical, conceptual models of trust in autonomous systems
- Human requirements for trust and trust in machines
- Machine requirements for trust and trust in humans
- Methods for engendering and measuring trust among humans and machines
- Metrics for deception among humans and machines
- Other computational and heuristic models of trust relationships, and related behaviors, in teams of humans and machines

Autonomy topics:

- Models of individual, group, and firm autonomous system behaviors
- Mathematical models of multitasking in a team (e.g., entropy levels overall and by individual agents, energy levels overall and by individual agents)

Network topics:

- Constructing, measuring, and assessing networks (e.g., the density of chat networks among human operators controlling multi-unmanned aerial vehicles)
- For networks, specify whether the application is for humans, machines, robots, or a combination, e.g., the density of inter-robot communications

After the symposium was completed, the book and the symposium took on separate lives. The following individuals were responsible for the proposal submitted to Springer after the symposium, for the divergence between the topics of the two, and for editing the book that has resulted.

Washington, DC, USA — Ranjeev Mittu
Washington, DC, USA — Donald Sofge
Atlanta, GA, USA — Alan Wagner
Augusta, GA, USA — W.F. Lawless

Contents

Chapter 1
Introduction

RanjeevMittu, Donald Sofge, AlanWagner, and W. F. Lawless

1.1 The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems

The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems addresses the current state-of-the-art in autonomy at the intersection of Robust Intelligence (RI) and trust, and the research gaps that must be overcome to enable the effective integration of autonomous and human systems. This is particularly true for the next generation of systems, which must scale to teams of autonomous platforms to better support their human operators and decision makers. This edited volume explores the intersection of RI and trust across multiple contexts among autonomous hybrid systems (where hybrids are arbitrary combinations of humans, machines and robots). To better understand the relationships between Artificial Intelligence (AI) and RI in a way that promotes trust between autonomous systems and human users, this edited volume explores the underlying theory, mathematics, computational models, and field applications.

To better understand and manage RI with AI in a manner that promotes trust in autonomous agents and teams, our interest is in the further development of theory, network models, mathematics, computational models, associations, and field

R. Mittu • D. Sofge
Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375, USA
e-mail: ranjeev.mittu@nrl.navy.mil; donald.sofge@nrl.navy.mil

A. Wagner
Georgia Tech Research Institute, 250 14th Street NW, Atlanta, GA 30318, USA
e-mail: Alan.Wagner@gtri.gatech.edu

W.F. Lawless (✉)
Paine College, 1235 15th Street, Augusta, GA 30901, USA
e-mail: WLawless@paine.edu

R. Mittu et al. (eds.), *Robust Intelligence and Trust in Autonomous Systems*,
DOI 10.1007/978-1-4899-7668-0_1

applications at the intersection of RI and trust. We are interested not only in effectiveness with a team's multitasking or in constructing RI networks and models, but in the efficiency and trust engendered among interacting participants.

Part of our symposium in 2014 sought a better understanding of the intersection of RI and trust for humans interacting with other humans and human groups (e.g., teams, firms, systems; also, the networks among these social objects). Our goal is to use this information with AI to not only model RI and trust, but also to predict outcomes from interactions between autonomous hybrid groups (e.g., hybrid teams in multitasking operations).

Systems that learn, adapt, and apply their experience to the problems faced in an environment may be better suited to respond to new and unexpected challenges. One could argue that such systems are "robust" to the prospect of a dynamic and occasionally unpredictable world. We expect the systems that exhibit this type of robustness to afford to those who interact with the system a greater degree of trust. For instance, an autonomous vehicle which, in addition to driving to different locations by itself, can also warn a passenger of locations where it should not drive, might likely be viewed as more robust than a similar system without such a warning capability. But would it be viewed as more trustworthy? This workshop endeavored to examine such questions that lay at the intersection of robust intelligence and trust. Problems such as these are particularly difficult because they imply situational variations that may be hard to define.

The focus of our workshop centered on how robust intelligence impacts trust in the system and how trust in the system makes it more or less robust. We explored approaches to RI and trust that included, among others, intelligent networks, intelligent agents, and multitasking by hybrid groups (i.e., arbitrary combinations of humans, machines and robots).

1.2 Background of the 2014 Symposium

Robust intelligence (RI) has not been easy to define. We proposed an approach to RI with artificial intelligence (AI) that may include, among other approaches, the science of intelligent networks, the generation of trust among intelligent agents, and multitasking among hybrid groups (humans, machines and robots). RI is the goal of several government projects to explore the intelligence as seen at the level of humans, including those directed by NSF (2013); the US Army (Army 2014) and the USAF (Gluck 2013). DARPA (2014) has a program on physical intelligence that is attempting to produce the first example of ""intelligent" behavior under thermodynamic pressure from their environment." Carnegie Mellon University (CMU 2014) has a program to build a robot that can execute "complex tasks in dangerous ... environments." IEEE (2014) has the journal *Intelligent Systems* to address various topics on intelligence in automation including trust; social computing; health; and, among others, coalitions that make the "effective use of limited resources to achieve complex and multiple objectives." From another

perspective, IBM has built a program that beat the reigning world champion at chess in 1997; another program that won at the game of Jeopardy in 2011 (du Sautoy 2014); and an intelligent operations center for the management of cities, transportation, and water (IBM 2014). Multiple other ways may exist to define or approach RI, and to measure it.

In an attempt to advance AI with a better understanding and management of RI, our interest is in the theory, network models, mathematics, computational models, associations, and field applications of RI. This means that we are interested in not only effectiveness with multitasking or in constructing RI networks and models, but in the efficiency and trust engendered among the participants during interactions.

Part of the goal in this symposium was to find a better understanding of RI and the autonomy it produces with humans interacting with other humans and human groups (e.g., teams, firms, systems; also, networks among these social objects). Our ultimate goal is to use this information with AI to not only model RI and autonomy, but also in the predictions of the outcomes from interactions between hybrid groups that interdependently generate networks and trust.

For multitasking with human teams and firms, interdependence is an important element in their RI: e.g., the Army is attempting to develop a robot that can produce "a set of intelligence-based capabilities sufficient to enable the teaming of autonomous systems with Soldiers" (Army 2014); and ONR is studying robust teamwork (ONR 2013). But a team's interdependence also introduces uncertainty, fundamentally impacting measurement (Lawless et al. 2013).

Unlike conventional computational models where agents act independently of neighbors, where, for example, a predator mathematically consumes its prey or not as a function of a random interaction process, interdependence means that agents dynamically respond to the bi-directional signals of actual or potential presence of other agents (e.g., in states poised to fight or flight), a significant increase over conventional modeling complexity; as an example of interdependence in Yellowstone's National Park (Hannibal 2012):

> aspen and other native vegetation, once decimated by overgrazing, are now growing up along the banks ... [in part] because elk and other browsing animals behave differently when wolves are around. Instead of eating down to the soil, they take a bite or two, look up to check for threats, and keep moving. [This means that the] greenery can grow tall enough to reproduce.

That the problem of interdependence remains unsolved, mathematically and conceptually, precludes hybrid teams based on artificial intelligence from processing information like human teams operating under interdependent challenges and perceived threats.

At this AAAI Symposium, we explored the various aspects and meanings of robust intelligence, networks and trust between humans, machines and robots in different contexts, and the social dynamics of networks and trust in teams or organizations composed of autonomous machines and robots working together with humans. We sought to identify and/or develop methods for structuring networks

and engendering trust between agents, to consider the static and dynamic aspects of behavior and relationships, and to propose metrics for measuring outcomes of interactions.

1.3 Contributed Chapters

Chapter 2 is titled "*Towards modeling the behavior of autonomous systems and humans for trusted operations.*" Its authors are Gavin Taylor, Ranjeev Mittu, Ciara Sibley and Joseph Coyne. The first author is with the U.S. Naval Academy; and authors Mittu, Sibley and Coyne are with the Naval Research Laboratory. In this chapter, the authors have studied the promise offered to the Department of Defense by autonomous robot and machine systems to improve its mission successes and to protect its valuable human users; but this promise has been countered by the increased complexity and workloads that have been placed on human supervisors by these systems. In this new era, as autonomy increases, the trust humans place in these systems becomes an important factor. Trust may depend on knowing whether anomalies exist in these systems; whether the anomalies that do exist can be managed; and whether these anomalies further affect the limitations of the human supervisors (acknowledged but not studied in this chapter). Using a mathematical manifold that captures a platform's trajectories to represent the tasks to be performed by an unassisted and unmanned autonomous system, the authors propose an example that exploits the errors generated for alarms and system analyses. The authors point out the existing research questions (e.g., user interaction patterns) and challenges that must also be addressed, including the best way for users to interact with autonomy; the optimized formal models of human decision-making; the modeling of active decision contexts; and the adaptation of concept drift techniques from the machine learning community.

Chapter 3 is titled "*Learning trustworthy behaviors using an inverse trust metric*"; its authors are Michael W. Floyd and Michael Drinkwater with Knexus Research Corporation, Springfield, Virginia; and David W. Aha, with the Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC. The authors present an algorithm by which a robot measures its own trustworthiness—an inverse trust metric—and uses this information to adapt its behavior, ideally becoming more trustworthy. They use case-based reasoning to gauge whether or not some previously used behavior or set of behaviors is likely to be trustworthy in the current environment. They present simulation experiments demonstrating the use of this "inverse trust metric" in a patrol scenario. The authors begin their chapter by assuming that those robots that can be trusted to perform important tasks may become helpful to human teams, but only if the humans can trust the robot as a member of a team to perform its assigned tasks as expected. But how to determine trust on the fly is a difficult problem. Instead of asking users how much they trust an autonomous agent, the authors use "inverse trust". They estimate the "inverse trust" for their concept as judged by the robot when determining its own

performance while working for multiple human operators, or in front of multiple human operators. Using simulation, the authors demonstrate the superiority of their case-based reasoning approach to random learning for robot team members assisting a human team. The authors conclude that more robotic uncertainty in performance impedes trust.

Chapter 4 is titled "*'Trust V'—Building and measuring trust in autonomous systems*"; its authors are Gari Palmer, Anne Selwyn and Dan Zwillinger with the Raytheon Corporation. The authors develop a framework based on the system V framework to codify how trust is built into a new system and how a system should respond in order to maintain trust. Their framework is a life-cycle model which adds trust components. Testing and evaluation ensure that the trust components are functional. During operational use these trust components allow the user to query the system to better understand (and trust) its operation. This framework is becoming more important as autonomous systems become more prominent; e.g., the Department of Defense has made autonomy one of its research priorities. Autonomous systems, those in use today and anticipated in the future, will need both system trust (i.e., when their specifications have been met) and operational trust (when the user's expectations have been met). Automated systems are more easily trusted than autonomous systems. But trusting complex automated systems requires rigorous Test & Evaluation (T&E) and Verification & Validation (V&V) processes. While similar processes are likely to be used to establish trust for autonomous systems, new methods set within these processes must address the unique attributes of autonomy, like adaptation to situations, or self-organization within situations. Using their framework, the authors identify specific methods for engendering trust in automated and autonomous systems, where systems range from automated to autonomous systems as endpoints. The authors give the example of a prototyped method that has been shown to enable trust. This framework supports the insertion of new methods to generate and measure operational trust in existing and future autonomous systems.

Chapter 5 is titled "*Big Data analytic paradigms—From principle component analysis to deep learning*"; its authors are Mo Jamshidi, Barney Tannahill and Arezou Moussavi with the Autonomous Control Engineering (ACE) Laboratory at The University of Texas, San Antonio (Tannahil is also from the Southwest Research Institute, or SwRI). This chapter presents an overview of Artificial Neural Networks (ANNs) ranging from multi-layer networks to recent advances related to deep architectures including auto-encoders and restricted Boltzmann machines (RBMs). Large sets of data (numerical, textural and image) have been accumulating at a rapid pace from multiple sources in all aspects of society. Advances in sensor technology, the Internet, social networks, wireless communication, and inexpensive memory have all contributed to the explosion of "Big Data" as this phenomenon has come to be known. Big Data is produced in many ways in today's interdependent global economy. Social networks, system of systems (SoS), and wireless systems are only some of the contributors to Big Data. Instead of a hindrance, many researchers have come to consider Big Data as a rich resource for future innovations in science, engineering, business and other potential applications. But the flood of data has

to be managed and controlled before useful information can be extracted. For the extraction of information to be useful, recent efforts have developed a promising approach known as "Data Analytics". This approach uses statistical and computational intelligence tools like principal component analysis (PCA), clustering, fuzzy logic, neuro-computing, evolutionary computation, Bayesian networks and other tools to reduce the size of Big Data. One of these tools, Deep Learning, is described by the development and use of neural networks in the machine learning community that has allowed for the extraordinary results recently obtained for digital speech, imagery, and natural language processing tasks. The authors present an example of Neural Networks using the data collected from a wind farm to demonstrate Data Analytics.

Chapter 6 is titled "*Artificial brain systems based on neural network discrete chaotic dynamics. Toward the development of conscious and creative robots*"; its author, Vladimir Gontar, prepared his chapter at the Biocircuits Institute, University of California in San Diego, while on a sabbatical; he has since returned to his affiliation at the Ben-Gurion University of the Negev in Israel. He is working on new theory and mathematical models of the human brain based on first principles for neural networks that model biochemical reactions to simulate consciousness. Consciousness is a hard problem. From Marcus and his colleagues (2014), although "no consensus" exists, current research tends to address how "systems might bridge from neuronal networks to symbolic cognition". Gontar's approach is similar. In contrast to regular information processes approximated linearly as a function of the energy available, he models information exchanges between neurons and neural networks based on the infinitesimally small energies needed to change chaotic systems. He compares the example of a mandala drawn by an artist matched step-by-step with one drawn by his chaos equations, concluding that this is how consciousness may be addressed computationally.

Chapter 7, on the "*Modeling and control of trust in human-robot collaborative manufacturing*", is authored by Behzad Sadrfaridpour, Hamed Saeidi, Jenny Burke, Kapil Madathil and Yue Wang with Clemson University and the Boeing Company. The authors explore trust in the context of Human-Robot Collaboration (HRC) on the factory floor. To measure and gauge the improvement in a system on the factory floor, they use a time-series model of trust, a model of a robot's performance to tie its speed to flexibility, and a model of a person that includes fatigue. They present a series of experiments which investigate how the robot and the human adapt to each other's changing performance and how these changes impact trust. HRC already exists on factory floors today, opening a new realm of manufacturing with robots in real-world settings. There, humans and robots work together by collaborating as coworkers. HRC plays a critical role in safety, productivity, and flexibility. Human-to-robot trust determines the human's acceptance and allocation of autonomy to a robot that in turn decides the efficiency of the task performed and the human's workload. Using Likert scales and time-series models of performance to measure trust, the authors studied trust in a robot in the laboratory subjectively and objectively under three control modes of the robot; viz., the robot placed under manual, autonomous and collaborative control conditions. Human operator control

was used in the manual condition; a neural network was used for intelligent control in the autonomous condition; and a mixed control was used in the collaborative condition. For this study, the authors did not find strong support for the autonomous mode. They also showed that under the collaborative mode, human-to-robot trust will be improved since the human has more control over the robot speed while the robot is adapting to the human speed as well.

Chapter 8 is titled "*Investigating human-robot trust in emergency scenarios: Methodological lessons learned*"; its authors Paul Robinette, Alan Wagner and Ayanna Howard are with the Georgia Institute of Technology; they conclude that trust has an elusive, subjective meaning depending on the context and the culture of the perceiver and the bias introduced by a questioner, especially in emergency scenarios. Being that few research protocols exist to study human-robot trust (HRT), the authors devised their own protocol to include risk on the part of both the human and the robot. Overall, they conclude that studies of HRT are inherently problematic, even though HRT has been studied as computational cognition; neurological change; and, among other studies, in the probability distributions of an agent's actions. They like Lee and See's claim that trust is an attitude associated with the goals sought under uncertainty and vulnerability. The authors performed experiments using crowdsourcing techniques. They found that the word phrasing of a narrative significantly affected decisions; that anchoring biases also had significant effects; and that unsuccessful robot leaders did not always dissuade their human followers. The latter finding presents a significant challenge to researchers to design robots in a way so that the robots communicate clearly with humans, so that humans do not overly-trust robots when they should not, and so that crowdsourcing for testing hypotheses provide generality and empirical evaluations if coupled with complementary methods (viz., narratives and simulated scenarios).

In Chap. 9, titled "*Designing for robust and effective teamwork in human-agent teams*", the authors Fei Gao, M.L. Cummings and Erin Solovey are with the Massachusetts Institute of Technology. The authors examine the impact of team structure, task uncertainty, and information-sharing tools on team coordination and performance. They present several information sharing tools which allow users to update others with regard to their status thus reducing work duplication and infrequent communication. The authors investigated the impact on human-agent teams of team structure, task uncertainty, and information-sharing, including coordination and performance. From their perspective, in the future, search and rescue, command and control, and air traffic control operators will be working in teams with robot teammates. But teams involve tasks that individual humans cannot do at all or are inefficient at doing. The authors contrasted organizational structures based on divisional teams, where self-contained redundancy governs under high uncertainty to make them more robust; and functional teams, where uncertainty is low and predictability is high. They discussed team situational awareness, where each member's contributions to and impacts on team tasks must be predictable and appreciated. The authors also discussed the costs of coordination and communication; and that these costs and duplication could be reduced with

information-sharing tools, while increasing robustness for divisional teams. The authors found that information sharing tools allowed users to communicate more effectively.

The author of Chap. 10, Kristin Schaefer, is with the U.S. Army Research Laboratory. In her article on "*Measuring Trust in Human Robot Interactions: Development of the 'Trust Perception Scale-HRI' "*, she studied the importance of trust in human-robot interaction and teaming as robotic technologies continue to improve their functional capability, robust intelligence, and autonomy. The author explores the development of a unifying survey scale to measure a human user's trust in a robotic system and in Human Robot Interaction (HRI) settings. She presents a series of related experiments leading to the creation of a 40 item survey which she argues measures trust across multiple conditions and robot domains. In this chapter, the author has summarized her PhD research to produce a reliable and valid subjective measure of the trust humans have of robots, the *Trust Perception Scale-HRI*. She performed an extensive literature review of trust in the interpersonal, automation and robot domains to determine if specific attributes accounted for human-robot trust. Schaefer developed an initial pool of items, tested it with human subjects, analyzed the results with a mental model of a robot, and reduced the number of items based on statistical and Subject Matter Expert (SME) content-validation procedures. This resulted in her 42 item scale plus a 14 item shorter scale derived from the feedback by her SMEs. She then used computer simulated human-robot interaction experimentation for a two-part task-based validation process to determine if the scale could measure a change in survey scores and measure the construct of trust. She first demonstrated that the scale measured a change pre-post interaction and across two reliability conditions (100% reliable feedback versus 25% reliable feedback) during a supervisory human-robot target detection task. This was followed by a second validation experiment using a Same-Trait approach during a team Soldier-robot navigation task. Her finalized 40-item scale performed well in both cases, and provided support for additional benefits when used in the HRI domain, above and beyond results achieved when using a previously developed automation-specific trust scale.

Chapter 11 is titled "*Methods for developing trust models for intelligent systems*"; its authors are Holly A. Yanco, Munjal Desai, Jill L. Drury and Aaron Steinfeld with the University of Massachusetts Lowell (Yanco and formerly Desai), The MITRE Corporation (Drury and Yanco), and Carnegie Mellon University (Steinfeld). The number of robots in use across the width of society, including in industry, with the military, and on the highways, is increasing rapidly, along with an expansion of their abilities to operate autonomously. Benefits from autonomy are increasing rapidly along with concerns about how well these systems can be, should be, and are being trusted. Human automation interaction (HAI) research is crucial to the further expansion of intelligence, but also its disuses and abuses. The research by the authors is designed to understand and model the factors that affect intelligent systems. The chapter begins with a review of prior research in the development of trust models, including surveys and experiments. Then the authors discuss two methods for investigating trust and creating trust models: surveys and robot studies.

They also produce 14 guidelines as well as an overall model of trust and the factors that increase and decrease trust. Finally, the authors review their conclusions and discuss the path forward.

In Chap. 12, titled, "*The intersection of robust intelligence and trust: Hybrid teams, firms & systems*", the authors are W.F. Lawless with Paine College and Donald Sofge with the Naval Research Laboratory; they are developing the physics of interdependent relations among social agents to reflect uncertainty arising from these relationships but also the power of social groups to solve difficult problems. Interdependence depends on the existence of alternative (bistable) interpretations of social reality. Interdependence makes social situations non-linear and non-intuitive, making interdependence a difficult problem to address. But if this problem can be solved, unlike today when robots work as individual agents, it will allow humans, machines and robots to work together in teams by multitasking to solve problems that only human teams can now solve. On the other hand, as interdependence increases across a group, its chances increase that it can make a mistake. Traditional models of interdependence consist primarily of traditional game theory. But game theory's solution of this problem relies heavily on increasing cooperation, thereby increasing static interdependence, further increasing the likelihood of a mistake. To avoid mistakes, the authors argue for a competitive situation similar to a Nash equilibrium, where the two sides engage in a nonlinear competition for neutrals (independent agents) to determine the winning argument at one point in time; mathematically, the result is a limit cycle as one side wins, but then that side falls behind in the next argument when the limits to its "solution" become apparent. The result is a method that increases social welfare. The authors describe how this may work in human-machine-robot environments.

References

Army (2014) Army robotics researchers look for into the future, D. Mcdally, RDECOM public affair, from http://www.army.mil/molule/article/?p=137837

CMU (2014) http://www.darpa.mil/Our_Work/TTO/Programs/DARPA_Robotics_Challenge/Track_A_Participants.aspx

Darpa (2014) http://www.darpa.mil/Our_Work/DSO/Programs/Physical_Intelligence.aspx

du Sautoy M (2014) The guardian (2012, 3/31), "AI robot: how machine intelligence is evolving". http://www-03.ibm.com/software/products/us/en/intelligent-operations-center/

Gluck K (2013) Robust decision making in integrated human-machine systems, U.S. Air Force BAA: 13.15.12.B0909

Hannibal ME (2012) Why the beaver should thank the wolf, New York Times. http://www.nytimes.com/2012/09/29/opinion/the-world-needs-wolves.html

IBM (2014) Smarter Cities. http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/

IEEE (2014) http://www.computer.org/csdl/mags/ex/2013/01/index.html

Lawless WF, Llinas J, Mittu R, Sofge DA, Sibley C, Coyne J, Russell S (2013) Robust Intelligence (RI) under uncertainty: mathematical and conceptual foundations of autonomous hybrid (human-machine-robot) teams, organizations and systems. Struct Dyn 6(2)

Marcus G, Marblestone A, Dean T (2014) The atoms of neural computation. Science 346:551–552

NSF (2013) Robust Intelligence (RI); National Science Foundation. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503305&org=IIS

ONR (2013) Command Decision Making (CDM) & Hybrid Human Computer Systems (HHCS) Annual Program Review, Naval Research Lab, Washington, DC, 4–7 June 2013

Chapter 2
Towards Modeling the Behavior of Autonomous Systems and Humans for Trusted Operations

Gavin Taylor, Ranjeev Mittu, Ciara Sibley, and Joseph Coyne

2.1 Introduction

Unmanned systems will perform an increasing number of missions in the future, reducing the risk to humans, while increasing their capabilities. The direction for these systems is clear, as a number of Department of Defense roadmaps call for increasing levels of autonomy to invert the current ratio of multiple operators to a single system (Winnefeld and Kendall 2011). This shift will require a substantial increase in unmanned system autonomy and will transform the operator's role from actively controlling elements of a single platform to supervising multiple complex autonomous systems. This future vision will also require the autonomous system to monitor the human operator's performance and intentions under different tasking and operational contexts, in order to understand how she is influencing the overall mission performance.

Successful collaboration with autonomy will necessitate that humans properly calibrate their trust and reliance on systems. Correctly determining reliability of a system will be critical in this future vision since automation bias, or overreliance on a system, can lead to complacency which in turn can cause errors of omission and commission (Cummings 2004). On the other hand, miscalibrated alert thresholds and criterion response settings can cause frequent alerts and interruptions (high false alarm rates), which can cause humans to lose trust and underutilize a system (i.e., ignore system alerts) (Parasuraman and Riley 1997). Hence, it is imperative that not only does the human have a model of normal system behavior in different

G. Taylor (✉)
United States Naval Academy, Annapolis, MD, USA
e-mail: taylor@usna.edu

R. Mittu • C. Sibley • J. Coyne
Naval Research Laboratory, Washington, DC, USA
e-mail: ranjeev.mittu@nrl.navy.mil; ciara.sibley@nrl.navy.mil; joseph.coyne@nrl.navy.mil

R. Mittu et al. (eds.), *Robust Intelligence and Trust in Autonomous Systems*,
DOI 10.1007/978-1-4899-7668-0_2

contexts, but that the system has a model of the capabilities and limitations of the human. The autonomy should not only fail transparently so that the human knows when to assist, but autonomy should also predict when the human is likely to fail and be able to provide assistance. The addition of more unmanned assets with multi-mission capabilities will increase operator demands and may challenge the operator's workload just to maintain situation awareness. Autonomy that monitors system (including human) behavior and alerts users to anomalies, however, should decrease the task load on the human and support them in the role of supervisor.

Noninvasive techniques to monitor a supervisor's state and workload (Fong et al. 2011; Sibley et al. 2011) would provide the autonomous systems with information about the user's capabilities and limitations in a given context, which could provide better prescriptions for how to interact with the user. However, many approaches to workload issues have been based on engineering new forms of autonomy assuming that the role of the human will be minimized. For the foreseeable future, however, the human will have at least a supervisory role within the system; rather than minimizing the actions of the human and automating those actions the human can already do well, it would be more efficient to develop a supervisory control paradigm that embraces the human as an agent within the system and leverages on her capabilities and minimizes the impact of her limitations.

In order to best develop techniques for identifying anomalous behaviors associated with the complex human-autonomous system, models of normal behaviors must be developed. For the purpose of this paper, an anomaly is not just a statistical outlier, but rather a deviation that prevents mission goals from being met, dependent on the context. Such system models may be based on, for example, mission outcome measures such as objective measures of successful mission outcomes with the corresponding behaviors of the system. Normalcy models can be used to detect whether events or state variables are anomalous, i.e., probability of a mission outcome measure that does not meet a key performance parameter or other metric.

The anomalous behavior of complex autonomous systems may be composed of internal states and relationships that are defined by platform kinematics, health and status, cyber phenomena and the effects caused by human interaction and control. Once the occurrence and relationships between abnormal behaviors in a given context can be established and predicted, our hypothesis is that the operational bounds of the system can be better understood. This enhanced understanding will provide transparency about the system performance to the user to enable trust to be properly calibrated with the system, making the prescriptions for human interaction that follow to become more relevant and effective during emergency procedures.

A key aspect of using normalcy models for detecting abnormal behaviors is the notion of context; and behaviors should be understood in the context in which they occur. In order to limit the false alarms, effectively integrating context is a critical first step. Normalcy models must be developed for each context of a mission, and used to identify potential deviations to determine whether such deviations are anomalous (i.e., impact mission success). Proper trust calibration would be assisted through the development of technology that provides the user with transparency

about system behavior. This technology will provide the user with information about how the system is likely to behave in different contexts and how the user should best respond.

We present an approach for modeling anomalies in complex system behavior; we do not address modeling human limitations and capabilities in this paper, but recognize that this is equally important in the development of trust in collaborative human-automation systems.

2.2 Understanding the Value of Context

The role of context is not only important when dealing with the behavior of autonomous systems, but also quite important in other areas of command and control. Today's warfighters operate in a highly dynamic world with a high degree of uncertainty, compounded by competing demands. Timely and effective decision making in this environment is challenging. The phrase "too much data—not enough information" is a common complaint in most Naval operational domains. Finding and integrating decision-relevant information (vice simply data) is difficult. Mission and task context is often absent (at least in computable and accessible forms), or sparsely/poorly represented in most information systems. This limitation requires decision makers to mentally reconstruct or infer contextually relevant information through laborious and error-prone internal processes as they attempt to comprehend and act on data. Furthermore, decision makers may need to multi-task among competing and often conflicting mission objectives, further complicating the management of information and decision making.

Clearly, there is a need for advanced mechanisms for the timely extraction and presentation of data that has value and relevance to decisions for a given context. To put the issue of context in perspective, consider that nearly all national defense missions involve Decision Support Systems (DSS)—systems that aim to decrease the cycle time from the gathering of data to operational decisions. However, the proliferation of sensors and large data sets are overwhelming DSSs, as they lack the tools to efficiently process, store, analyze, and retrieve vast amounts of data. Additionally, these systems are relatively immature in helping users recognize and understand important contextual data or cues.

2.3 Context and the Complexity of Anomaly Detection

Understanding anomalous behaviors within the complex human-autonomous system requires an understanding of the context in which the behavior is occurring. Ultimately, when considering complex, autonomous systems comprised of multiple entities, the question is not what is wrong with a single element, but whether that anomaly affects performance of the team and whether it is possible to achieve the

mission goals in spite of that problem. For example, platform instability during high winds may be normal, whereas the same degree of instability during calm winds may be abnormal. Furthermore, what may appear as an explainable deviation may actually be a critical problem if that event causes the system to enter future states that prevent the satisfaction of a given objective function. The key distinction is that in certain settings, it may be appropriate to consider anomalies as those situations that effect outcomes, rather than just statistical outliers. In terms of the team, the question becomes which element should have to address the problem (the human or the autonomy).

The ability to identify and monitor anomalies in the complex human-autonomous system is a challenge, particularly as increasing levels of autonomy increase system complexity and, fundamentally, human interactions inject significant complexity via unpredictability into the overall system. Furthermore, anomaly detection within complex autonomous systems cannot ignore the dependencies between communication networks, kinematic behavior, and platform health and status.

Threats from adversaries, the environment, and even benign intent will need to be detected within the communications infrastructure, in order to understand its impact to the broader platform kinematics, health and status. Possible future scenarios might include cyber threats that take control of a platform in order to conduct malicious activity, which may cause unusual behavior in the other dimensions and corresponding states. The dependency on cyber networks means that a network provides unique and complete insight into mission operations. The existence of passive, active, and adversarial activities creates an ecosystem where "normal" or "abnormal" is dynamic, flexible, and evolving. The intrinsic nature of these activities results in challenges to anomaly detection methods that apply signatures or rules that have a high number of false positives. Furthermore, anomaly detection is difficult in large, multi-dimensional datasets and is affected by the "curse of dimensionality." Compounding this problem is the fact that human operators have limited time to deal with complex (cause and effect) and/or subtle ("slow and low") anomalies, while monitoring the information from sensors, and concurrently conducting mission planning tasks. The reality is that in future military environments, fewer operators due to reduced manning may make matters worse, particularly if the system is reliant on the human to resolve all anomalies!

Below we describe research efforts underway in the area of anomaly detection via manifolds and reinforcement learning.

2.3.1 Manifolds for Anomaly Detection

A fundamental challenge in anomaly detection is the need for appropriate metrics to distinguish between normal and abnormal behaviors. This is especially true when one deals with nonlinear dynamic systems where the data generated contains highly nonlinear relationships for which Euclidean metrics aren't appropriate.

One approach is to employ a nonlinear "space" called a manifold to capture the data, and then use the natural nonlinear metric on the manifold, in particular the Riemannian metric, to define distances among different behaviors.

We view the path of an unmanned system as a continuous trajectory on the manifold and recognize any deviations due to human inputs, environmental impacts, etc. Mathematically, we transform the different data types into a common manifold-valued data so that comparisons can be made with regard to behaviors.

For example, a manifold for an unmanned system could be 12 dimensional, composed of position, pitch, roll, yaw, velocities of the position coordinates, and angular velocities of the pitch, roll and yaw. This 12-dimensional model captures any platform (in fact any moving rigid object's) trajectories under all possible environment conditions or behaviors. This manifold is the tangent bundle, TM of $SO(3) \times \Re^3$. Here $SO(3)$ denotes the set of all possible rotations of the unmanned system which is a Lie group, and $\Re^3$ the set of all translations of the platform. Since rotations and translations do not commute, this is not a direct product of $SO(3)$ with $\Re^3$. The product between $SO(3)$ and $\Re^3$ is a "Semi-Product" $\ltimes$. Non-linear key geometric, dynamical and kinematic characteristics are represented using TM. This manifold model is able to encapsulate the unique structure of the environment, effects of human behaviors, etc. through continuous parameterizations and coherent relationships.

Once we have this manifold model and its Riemannian metric, it is possible to define concepts of geodesic neighborhood and other appropriate measurements and map those to mission cost. Such a mapping is done by designing a weighted cost function with dynamical neighborhoods around a trajectory of the platform. For example, if the weather is good in the morning, the neighborhood is smaller than it would be with bad weather. This innovative manifold method could be used to dynamically identify normal or abnormal behaviors occurring during a mission, taking into consideration whether a mission could be successfully achieved under a given cost constraint. We also have the freedom to adjust normal neighborhoods if a mission suddenly changes while en-route. Our model is robust and captures complicated dynamics of unmanned systems and is able to encapsulate very high dimensional data using only a 12 dimensional configuration space.

The algorithms use continuous parameterizations and coherent relationships and are scalable. Our manifold-based methods provide new techniques to combine qualitative (platform mechanics) and quantitative (measured data) methods and are able to handle large, nonlinear dynamic data sets.

2.4 Reinforcement Learning for Anomaly Detection

The military and commercial communities increasingly rely on autonomous systems to augment their capabilities. For example, power plants feature automatic monitoring and safety features, and the military increasingly employs unmanned systems in denied or politically sensitive theaters. However, it is rare for these