Human-AI Decision Systems

November 1, 2017
Human-AI Decision Systems
Alex (Sandy) Pentland
MIT Media Laboratory, Cambridge, MA
pentland@mit.edu

DRAFT 11/1/2017

Abstract. We propose three concepts to guide the creation of high performance human-AI decision systems. These concepts are inspired by successes within the commercial sector and academia that successfully integrate information across many domains. The first concept is a framework for integration of AI capabilities into the enterprise that optimizes trust and performance within the workforce. The second is an approach for facilitating multi-domain operations though real-time creation of multi-domain ``task teams'' by dynamic management of information abstraction, teaming, and risk control. Lastly, we describe a new paradigm for multi-level data security and multi-organization data sharing that will be a key enabler of AI-enhanced multi-domain operation in the future.

Keywords: Human-AI decision systems, Multi-domain decision systems, distributed Bayesian portfolio analysis, trust networks

1. Introduction

Perhaps the most critical function of any organization or society is its decision systems. In modern societies, decision systems provide a leader the ability to make informed and timely decisions, supported by a complex enterprise of distributed information and communication systems that provide situational awareness. Traditionally, decision systems have been confined to individual physical domains, such as logistics, physical plant, and human resources, and more recently virtual domains such as cyber, by both policy and technology, resulting in challenges with the integration of information across disparate domains.

However, the complexities of modern peer competition necessitate the ability to integrate and operate concurrently and jointly across multiple domains. Migrating to a multi-domain operating construct has many challenges, across technology, manpower, doctrine, and culture. Existing decision systems are built on legacy hardware and software using proprietary data analysis and exchange mechanisms, inhibiting modernization and wider integration. Human- and sensor-generated data from a wide-range of sources and organizations are combined in ad-hoc ways and stored and compartmentalized at multiple security levels on multiple networks, often inhibiting joint multi-organization information sharing and decision-making. Additionally, these systems have little automation and are human-intensive to operate and will not scale as manpower remains fixed and as mission scope and responsibility increases.

Even where automation and analytics do exist, they suffer from disuse and poor trust calibration due to lack of training, transparency, and measured performance, both real and perceived. Meanwhile, there is widespread acknowledgement that future strategic advantage depends on the ability to leverage Artificial Intelligence (AI), such as machine learning, computer vision, and autonomous systems, and integrate it with the workforce to create symbiotic human-machine teams.

In this paper, we propose three concepts to address these and other challenges in realizing an integrated multi-domain operating construct, and compare and contrast similar approaches that have been proposed. First, we must develop a framework to aid the development, maturation, and diffusion of AI capabilities into the enterprise and institutionalize processes that optimize trust and performance within the workforce. Second, we explore how to enable the creation of multi-domain human operators through balanced information management and dynamically managed risk. Lastly, we describe a new paradigm for multi-level data security and multi-organization data sharing that will be a key enabler of joint and coalition multi-domain operation in the future.

1.1 Human-compatible AI framework for dynamic organizations

Today there is a deep mistrust of user-facing automation and automatic AI systems. As a consequence, capabilities that can reduce the human-intensive nature of operations go unused, and analysts and operators often engage in manual unassisted tradecraft because it is what they know and understand. Often this distrust is justified because of performance issues; lack of transparency and agency; and poor user trust calibration with the AI system.

Performance issues can arise from a multitude of areas including the choice of algorithm, data fidelity, and implementation. However, one of the most frequent problems arise when AI systems use out-of-domain data for initial training and are never retrained once deployed in-domain. A common example is natural-language processing capabilities, such as statistical named entity recognition and task-oriented information, that are frequently trained on large corpora of well-written English news reporting (out-of-domain), but yield poor

performance when applied against jargon filled local sources (in-domain). Additionally, many capabilities are trained in bulk increments, which lose performance over time as data evolves, and do not employ active learning methods that provide the AI systems with the ability to elicit feedback from users to improve the underlying models over time.

Next, AI systems are often not used because they are built as black boxes and do not provide transparency into what assumptions and decisions the underlying algorithms are making on the user's behalf. For humans to effectively team with automation, displays are needed that can explain and visualize what decision criteria the system uses. Recent studies suggest that this transparency can improve user trust and automation adoption [b6]. Additionally, employment of these technologies comes with few user-configurable options, resulting in a one-size-fits-none system that can stifle user agency. For example, forecasting algorithms are a class of capabilities that can have been shown to yield greater performance than human-only decision-making, yet users traditionally will not use them because they know they are imperfect. However, when given the ability to change even a single algorithmic parameter and inspect how the output changes in response, the trust and usage of the forecasting aid can increase [b7].

Performance of the joint human-AI system is also impacted by how much a user relies on automation and if that reliance is properly calibrated. Calibrated trust is achieved when a user's reliance in an AI system matches the trustworthiness (capability) of that system. However, in real-world implementations trust calibration often ends up near the extremes of this spectrum, specifically ``overtrust''; when a user's trust exceeds the system's capabilities, leading to misuse; and ``distrust'', when trust falls short of the system's actual capabilities, leading to disuse [b8].

But there is another path forward: rather than attempting to replace human operators by fully autonomous machines, it has been shown that certain kinds of human-AI combinations will perform better than humans and AI working alone. Although no person is better than a machine for many repetitive support tasks or focused tactical problems, no machine is better than a man with a machine for difficult tasks such as analysis and interpretation [b4]. Thus, by building AI systems that are compatible with human behavior, and specifically AI systems which leverage the manner in which humans use social information, we can build human-AI decision systems that extend human intelligence capabilities.

1.2 Organizational Shifts from Static to Dynamic

Over the last century, management systems have evolved from static to dynamic and from siloed to wholistic. For instance, consider the first consumer-facing financial investment institutions. They attracted investors with just one or two investment strategies, which varied little over time. Similarly, the first department stores competed by offering a wide but static range of products. Such businesses thought of strategy and tactics as fixed, and had very little investment in discovering new products or assessing potential customer demand. And until very recently, the Armed Forces were similar static organizations. Historical strategies, tactics, and organizational patterns were taught to new officers, and much of the organization of war would be recognizable to the grandfathers of the last generation of warfighters.

A cycle of plan-execute-assess, using a fixed array of intelligence sources, action capabilities, and organizational plans, is the signature of competition in a static world. It does not work in a world that is changing quickly or against more agile competitors. For instance, the financial world is now dominated by hedge funds that are continuously testing new strategies using machine learning and stochastic analysis techniques, and continuously seeking new information sources. The average lifetime of entire families of investment strategies is now under six months, and new information sources are being continuously evaluated. The department store has been replaced by Amazon and similar agile markets, who have a huge range of products that are continuously evaluated. Live testing of several different sales strategies are launched each and every day, and extensive, continuous stochastic modeling of customer behavior provides for next-day delivery of most products anywhere in the country while simultaneously reducing fraud by an order of magnitude as compared to more static-minded competitors.

1.3 Limitations of Reinforcement Learning

A similar evolution has happened in the AI technology that has been incorporated into competition systems. Beginning with alpha-beta ``smart search'' for making decisions in competitive games such as chess, the technology has moved to competitive techniques that allow machines to learn which strategies are most effective. These approaches are often called reinforcement learning, and Q-learning. Additionally, many of these modern approaches to artificial intelligence make use of *deep learning*, which uses larger, more complex neural network algorithms than were practical a few years ago, leading to more complex behaviors that can be modeled by a machine. These approaches search a space of game strategies and learn the strengths and weaknesses of competitive systems by playing simulated games that pit one deep learner against another. This class of technique is now competitive in hugely complex games such as Go. Although the achievement of Go-playing computers is impressive, note that this approach is fundamentally limited in several ways: there is a fixed menu of actions, the goals are clear and unidimensional, the game situation is known precisely and instantly, and the space of strategies used by human opponents is relatively limited.

Because of the fixed menu of actions, clear numerical goals, and a well-behaved strategy space, the current methods of AI, such as deep learning, can be used successfully. Deep learning and stochastic methods, such as Random Forest, model the competitive strategy space by piecewise linear ``chunks'' that neither capture the continuous nature of the strategy space (or physical world) nor align with the cognitive space used by humans to understand and act. These problems of current AI methods are hidden by smoothness constraints that are implicit in the large numbers of human-generated tactical examples that are required to train the network. In essence, the deep Q-learning

approach creates a Frankenstein-like model of the human decision tree using linear fragments. Just as in computer graphics, with enough linear approximations, you can successfully model continuous functions.

This means that the same methodology cannot easily be used in complex, real-world operations. The central problem is that real operations are not stationary, that is, the payoff for an action changes over time. This evolving payoff makes it difficult to adapt today's AI methods and will lead to systems that will fail to yield good results, and these failures will be unexpected and catastrophic. Similarly, their reliance on brute-force piece-wise linear approximation produces models which are confusing and anti-intuitive for human participants. In addition, there is usually much less training data available in military problems, the menu of actions is both much larger and continuous, and knowledge of the situation is both partial and stochastic.

2. Human Compatible AI Systems: A Framework

The insight powering today's cutting-edge artificial intelligence algorithms, including both the deep-learning/neural-network approaches and statistical machine learning algorithms like Random Forests, is that almost any function can be modeled by using a complex network of many simple logic machines. In these AI methods, the connections between simple logic machines ("neurons") are reconfigured as the system learns, with each training example slightly changing the connections between all of the neurons based on how the connection contributes to the overall outcome.

Most of the recent progress in AI techniques has been due to better solutions to the *credit assignment problem*, which is the algorithm that determines how much each connection contributed to the overall answer. Once the contributions of each connection have been determined, then the learning algorithm that builds the AI is simple: connections which have contributed positively are strengthened; those that have contributed negatively are weakened. The credit assignment problem is a long-standing problem in AI research and there are no easy answers. Better solutions to the credit assignment problem for human-machine organizations will be one of the key research areas for human-AI organizations in the 21st century and one of the areas that will be the most fruitful for managing organizations.

This same insight can be used to build a human-AI ecosystem. Imagine a human organization as a kind of brain, with humans as the individual neurons. Static firms, symbolized by the ubiquitous org chart, have fixed connections and as a result, a limited ability to learn and adapt. Typically their departments become siloed, with little communication between them so that the flow of fresh, cross-cutting ideas is blocked. As a consequence, these statically-configured, minimally interconnected organizations risk falling to newer, less ossified competitors.

But if an organization's skills can be supercharged by adopting the right sort of credit assignment function, then the connections, among individuals, teams, and teams of teams, might continuously reorganize themselves in response to shifting circumstances and challenges. Obviously, this continuously adapting and learning organization, even if assisted by sophisticated algorithms, will require organizations to adapt policy, training, doctrine and a wide spectrum of other issues. The potential gain is the emergence of fluid organizations that have a faster orient-observe-decide-act (OODA) loop than a potential competitor.

The AI in such a smart organization would be used to create the best connections between people and ideas, not for replacing the neurons (people) within a static, frozen organization.

Instead of people being trained to be simple rule-following machines (or replace them by AIs), people would be trained to engage in continuous improvement that has been characteristic of the *Kaizen-style* manufacturing of Toyota. Similarly, on-line marketers such as Amazon and Google, and financial services firms such as Blackrock and Renaissance, have started to adopt these approaches. Such dynamic organizations principally use AI to connect people to information and to other people's projects, not to just replace people.

Because humans have more general capabilities than simple logic machines, such a fluid organizational architecture can be qualitatively more powerful that today's AIs. Armed with the right credit assignment feedback, human *smart neurons* in an organizational brain can fill communication gaps to accelerate learning, anticipate unknowns and invent new structures to address emerging market forces and trends. It can be more resilient and powerful than an organization where people have been replaced by AI modules.

2.1 Mathematical Framework for Human Compatible AI Systems

How can we achieve this vision of a dynamic, human-AI compatible organization? The key is to have a credit assignment function (a reward function) that makes sense for each individual and yet at the same time yields global optimal performance. The phrase *makes sense* means that each individual must be able to easily understand their options and how to make choices that are good for them as an individual as well as good for the whole organization. They must be able to easily and clearly communicate the choices they have made and the anticipated or actual outcomes resulting from those choices. Only through this sort of understanding and alignment of incentives within an enterprise will the human participants come to trust both the AI systems and the organization. As noted previously, it is the development of the credit assignment function and an adaptive organization to capitalize upon it that will be key to building human-AI organizations for the future.

Fortunately, in the last two decades, analytical modeling methods have been developed that solve the problem of continual strategic improvement under precisely these conditions. These methods do not suffer from the shortcomings of other current AI methods; they handle changeable, non-stationary situations extremely well, produce models that are intuitive to humans, and are competitive with the

best deep learning methods. They also apply directly to situations with stochastic and incomplete data and continuous high-dimensional action spaces.

This family of AI methods is often referred to as distributed Bayesian portfolio analysis, which is a generalization of the well-known Thompson sampling methods first discovered in the 1930s. These methods are used to choose among alternative actions when the associated utilities are unknown or uncertain [b5]. Both the best hedge funds and the best consumer retail organizations use this type of AI method for their overall management structure, though these organizations often do not call it by this name. The key generalizations that these portfolio methods have in common with standard Thompson sampling is that the evaluation of alternative choices depends both on the action space and the information sources, and that many strategy families are explored simultaneously, rather like simulated annealing methods.

The core idea associated with these analysis methods and Thompson sampling is that when decision makers are faced with a wide range of alternative actions, each with unknown payoff, they have to select actions to discover those that lead to the best payoffs, and at the same time exploit the actions that are currently believed to be the best in order to remain competitive against opponents. These methods are designed so that the same decision logic applies both to directly competitive actions and to information gathering actions. As a decision maker discovers the actions that have the best payoffs, or produce the best information, they apply Bayes rule to simultaneously sample the various available alternatives. In this case, Bayes rule can be thought of as the likelihood of an outcome times a quality assessment. In this distributed framework, the prior likelihood can be efficiently determined by observing the payoffs of other members of a decision maker's distributed team. This use of social learning dramatically improves both overall performance and reduces the cognitive load placed on the human participants. The ability to rapidly communicate and observe other decisions across the enterprise is one of the key aspects of optimal human-AI organizations.

It is important to emphasize that this approach is qualitatively the same as that used by Amazon to configure its portfolio of products as well as its delivery services. A very similar approach is taken by the best financial hedge funds. Fully dynamic and interleaved planning, intelligence gathering, evaluation, and action selection produce a powerfully optimized force.

This Bayesian AI-portfolio approach has one more advantage that is absolutely unique and essential for systems that combine humans and AIs: the actions of the individual human are both in their best interest and in the best interest of the organization. Furthermore, the alignment of individuals' incentives and the organization's incentives are visible and understandable.

2.2 The Human Experience in a Human-AI Organization

What does such a human-AI organization feel like to human participants? It turns out that high performing teams naturally behave in exactly the manner required for a successful human-AI organization. Using cell phones with tracking software and similar technical means, MIT Connection Science researchers [b9] have characterized the patterns of behavior that associate with high performance. A series of experiments were run in 2010 in which standard IQ tests and other measures were administered to nearly 700 people. The study participants were next divided into teams of two to five members and then given a variety of problems to solve.

Somewhat surprisingly, it was found that a group's success at meeting these challenges was only weakly related to the IQs of its individual members. So, too, little correlation was found with the group's cohesion or levels of motivation and satisfaction, as measured with standardized questionnaires. Instead, the most successful teams were those that were able to optimize communication within the group. If every team member was engaged and making many contributions, then the group was very likely to be successful. This also meant that members whose ideas and experience were different from the majority had the opportunity to contribute and be heard.

In a follow-up study, the research team was able to show that the same strong pattern of exchanges that gives rise to successful teams also produces what has been described as a “team of teams,” and which is called “X-Teams” here at MIT [b10]. These emergent communication structures form meta-teams---groups that assemble collaborators from teams spread across different parts of the organization---that help spread innovation throughout the organization. The research results demonstrate that people who are especially adept at finding and maintaining connections across an organization are critical for opening up the channels needed to spread ideas more broadly across an organization. These cross-team ties help to break down silos and increase an organization's productivity and ability to innovate.

The distributed portfolio AI methods provide the quantitative guidance required to shape the strategic communications architecture and thus optimize total organizational performance. In other words, it is these distributed AI methods that provide the continual feedback and optimization that transforms a traditional human organization into a human-AI symbiosis. The key to this transformation is that the guidance given by the AI is in the direct best interest of each individual and simultaneously in the best interest of the organization, so that the incentives of the individual and those of the organization are aligned. Moreover, this is a familiar framework for humans -- indeed, the evidence is that this type of framework is built into the genes of all social species, and it is this alignment that enables the existence of social species that reproduce individually.

As an example, the **MIT Connection Science (http://connection.mit.edu)** research team recently examined how top performing teams and teams-of-teams maximize the sharing of strategic information within a social-network stock-trading site where people can see the strategies that other people choose, discuss them, and copy them. The team analyzed some 5.8 million transactions and found that the groups of traders who fared the best all followed the distributed Thompson sampling methodology. It was calculated that the forecasts from groups that followed the distributed Thompson sampling formula reliably beat the best individual forecasts by a margin of almost 30

percent. Furthermore, when the results of these groups were compared to results obtained using standard AI techniques, the humans that followed the distributed Thompson sampling methodology reliably beat the standard AI techniques.

3.3 First Steps Toward a Human-AI Organization

How can senior leaders begin to implement a human-AI organization? The first steps should be focused on real-time communication of the successes and failures that other teams have experienced from implementing strategies and tactics appropriate to the action capabilities of the team. Even co-located workers have a limited capacity to know everything that is happening in the local neighborhood of an organization. To remedy this problem, MIT researchers have used standard AI algorithms to generate a prioritized list of potentially productive new connections within a larger organization. These algorithms are able to sort through gigabytes of action-process data to communicate the strategy and tactics used by people with similar responsibilities and situations in different divisions. Critically, the distributed AI portfolio methods can also compare actual patterns of communication with what theory predicts is optimal in order to check for communication gaps and inefficiencies.

Actively encouraging greater engagement among team members offers yet another mission-critical benefit: when everyone participates and shares ideas, individuals feel more positive about belonging to a team, and develop greater trust in their colleagues. These feelings are essential for building organizational resilience. Social psychology has documented the incredible power of group identities to bond people and shape their behavior and the same holds true in the workplace; group membership provides the social capital needed to see team members through inevitable conflicts and difficult periods.

An important lesson from the hundreds of case studies that the **MIT Connection Science (http://connection.mit.edu)** team has conducted is that the decisions of top leaders can consistently improve by using these AI portfolio techniques to better incorporate the experiences of employees who actually have skin in the game. For instance, front-line workers often have better ideas about how to deal with challenging situations than the managers, and tactical engineers know more about how a new capability is shaping up than its designers do. The secret to creating an agile, robust organization is closing the communications gap between workers and bosses so that employees are both helping to create corporate plans and executing them. This closure fits with another key finding: developing the best strategy in any scenario involves striking a balance between engaging with familiar practices and exploring fresh ideas.

3. Creating cross-domain operations and integrated decision-making

To effectively migrate to an AI-enhanced cross-domain geo-intelligence operations, changes to doctrine, organization, training, and technology are required in order to create native cross-domain operators and decision-makers. These native operators' focus would not be organized around the concept of individual domains or data resources, but instead on the complex condition-action pairings that span multiple domains in order to achieve integrated effects.

In a native cross-domain framework, the operator would have real-time integrated information at the right level of the organization in order to detect and diagnose problems, along with the ability and authority to dynamically reconfigure available assets.

3.1 Cross-Domain Operation through Task-Oriented Information Management

While the promise of cross-domain operators is clear, the instantiation could be achieved by two competing approaches: the human-only approach, which relies on extensive training in specific domains to develop deep subject matter experts; or the human-assisted approach, where technology is used to abstract information, manage risk, more dynamically organize stakeholders.

In the human-assisted approach, technology is used to mediate what is exposed to the cross-domain operator and distill data so as to provide the minimum set of information required to maintain awareness and make informed decisions. This concept is not new and is how complex sociotechnical, real-time, and dynamic systems in fields such as nuclear power, aviation, and medicine have been designed and managed.

For example, even very experienced nuclear control room operators do not have a complete understanding of how the entire complex system (fission reactors, steam turbines, cooling systems) works. To achieve this tasked-oriented information management, processes like ecological interface design are often used which focus on the work domain and environment, rather than on a specific task or end-user.

This design methodology is based on two concepts from cognitive engineering; the causal information abstraction hierarchy, which governs how information should be presented; and the Skills, Rules, Knowledge taxonomy, which defines the psychological processes of how humans process information. These frameworks and others will be required to carefully study and design the trade-offs between the operational efficiencies gained and risk induced from the use of abstracted information within the organization.

3.2 Rapid Response Dynamic Teaming

With the realization of a cross-domain decision system that can provide timely and effective situation understanding, the operator now needs the ability to find and connect with the right people in the enterprise once he or she know something requires greater attention. This process is similar to the rapid response teams used in hospitals, such as when a nurse knows something is not quite right with a patient, they have the ability and authority to immediately mobilize 2--3 additional sets of eyes to the bedside so that the just-formed team can quickly triage and determine an appropriate course of action for the patient. This teaming process works because the hospital is essentially a single system and the number of unique organizational roles is relatively small. However, most large organizations are more akin to a system-of-systems, and it is likely not known what roles are required to form a rapid response team to respond to cross-domain issues.

To address this challenge, we can take advantage of ``smart" communications systems that use AI to find and form teams of the correct people at the correct time. This AI system would monitor the status of many complex systems and operational conditions, and process structured organizational data to determine and connect a rapid response team. Then, like the health care model, the assembled actors would have the authority and tactical freedom to dynamically select and execute actions using any combination of their capabilities. In exchange for this operational freedom, the system would control real-time risk by continuously auditing the resources that have been used or at risk, and calculating the opportunity costs for not addressing deferred tasks. This approach should provide the organizational scale and flexibility required for cross-domain decisions and also the agility to interoperate at the speed of competition in the future.

4. Cryptographic trust networks for secure data sharing

It is widely believed that most future operations will be collaborations between several organizations. A significant challenge with these missions is data sharing in level security environment. Cryptographic trust networks have the potential to address a number of challenges with data sharing in these environments. Current research in trust networks and personal secure data stores are outlined in [b11], [b12], [b13], [b14], and [b15].

4.1 Multi-Organization Multi-Level Data Sharing Challenges

There are a number of different challenges with multi-organization, multi-security level data sharing. The first challenge is ensuring the availability of the relevant information that mission partners require to execute their tasks while preventing access to information that they are not authorized to receive. Part of the difficulty with data security is that the classification level of information can change as it flows through a processing chain. Recipients may have permission to receive processed information at a downstream point but not have access to information on upstream sources, means, and methods.

A second challenge is ensuring information integrity in large systems with complex data flows. Success relies on the assurance that data and algorithms have not been spoofed, corrupted, diverted, or destroyed at some point in the processing system. A capability is required to be able to verify that processing chains are secure. This will require that information workflows are continuously monitored to verify and validate the integrity of the data, applications, and supporting infrastructure.

There is a growing interest in computing architectures like cloud computing because of important capabilities that they can provide to an organization, like ubiquity of computation, on-demand scaling, user convenience, and improved reliability and resource utilization. An aspect of the vision of the new computing architectures is the ``as a service'' concept. The design of these architectures is driven to make software, platforms and infrastructure into services that are provided to users. Everything from applications, data storage, networks, and operating systems, are services. In a collaborative operation partners with compatible cloud computing resources, additional economies of scale, redundancy, and reliability can be achieved by pooling these resources. There are challenges to ensuring that this pooling can be accomplished while still ensuring data security and integrity across partner organizations, especially when storage and computation may be performed on computer systems of partners without permission to access the associated information. Crypographic trust networks have the potential to address all of these challenges.

4.1 A New Technical Approach

To build a successful human-AI organization requires retooling geo-intelligence protocols. To accomplish this in a secure, efficient manner requires retooling the digital infrastructure. Our OPen ALgorithms (OPAL) paradigm enables information exchange based on safe predefined queries that preserve security at all levels. The phrase “Open Algorithms” is drawn from best practice in designing cryptographic algorithms: because the stakeholders are able to examine the algorithms, it is extremely difficult for malicious actors to insert back door vulnerabilities.

Instead of requiring sensitive data to be shared, OPAL enables predetermined algorithms, that are consistently evaluated in a lightweight process by human stakeholders, to be sent to decentralized data stores to fetch the analysis results. For added security, the data within OPAL is always in an encrypted state, including during transmission and computation. OPAL represents a new paradigm of knowledge-sharing which fundamentally improves security while at the same time making access to critical knowledge more flexible and more efficient.

OPAL is enabled by the emergence of blockchain cryptography technologies that provide more secure forms of user-controlled data management. These are based on a decentralization of the authentication and authorization functions that are core to security management. A blockchain-based credential management and access control system can be based on security-maximizing algorithms that are authenticated by diverse sets of relationships.

A key idea of the OPAL architecture is that it should implement AI systems that support rather than supplant individual human autonomy, a requirement that strongly constrains the design of the underlying technical architecture. To accomplish this goal, OPAL is based on the idea of a *trust network*.

Trust networks possess a number of important criteria. They contain distributed, encrypted databases that run on different computer systems to provide defense in depth against cyberattacks from the outside. Any single hacker exploit should result in access to only a limited part of the entire database. Every data operation requires a reliable chain of identity and access credentials so stakeholders can know where data came from and where they went. All entities are subject to metadata monitoring and investigative auditing.

Trust networks combine a computer network that keeps track of user permissions for each piece of data within a legal framework that specifies what can and cannot be done with the data, and what happens if there is a violation of the permissions. By maintaining a tamper-

proof history of provenance and permissions, trust networks can be automatically audited to ensure that data usage agreements are being honored. Previously, trust networks were complex and expensive to run, but the decreasing cost of computing power has brought them within the reach of smaller organizations and even individuals. Trust networks have the potential to alleviate several problems for the Department of Defense related to insider threats. Specifically, the concentration of large amounts of information in the hands of one user, and those trusted users having almost complete control over that information.

The connection science research group at MIT has helped build OPAL (OPen ALgorithms), a national-scale version of a Trust Network, with support from MasterCard, Intuit, UBS, the French Government, and the World Bank. This system may be viewed as an extension of the Estonian X-Road system which has successfully resisted Russian cyberattacks for the last 15 years.

Bibliography

b3. R. Work (2015, December, 14) Center for a New American Security Defense Forum, [Online]. Available: **https://www.defense.gov/News/Speeches/Speech-View/Article/634214/\\cnas-defense-forum/**

b4. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Anita Williams Woolley et al. in Science, Vol. 330, pages 686–688; October 29, 2010

b5. Thompson, William R. ``On the likelihood that one unknown probability exceeds another in view of the evidence of two samples''. Biometrika, 25(3–4):285–294, 1933.

b6. Lyons J.B. et al., Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In: Savage-Knepshield P., Chen J. (eds) Advances in Human Factors in Robots and Unmanned Systems 2017

b7. Dietvorst, B.J., Simmons, J.P. and Massey, C., 2016. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science 2016

b8. Adapted from: Lee, John D., and Katrina A. See. ``Trust in automation: Designing for appropriate reliance.'' Human Factors: The Journal of the Human Factors and Ergonomics Society

46.1 (2004): 50-80.

b9. MIT Connection Science. **http://connection.mit.edu/ (http://connection.mit.edu/)**

b10. Ancona, D., Bresman, H., and Kaeufer, K., The Comparative Advantage of X-Teams, MIT Sloan Management Review, Spring 2002:33-39, 2002.

b11. de Montjoye, Yves-Alexandre, et al. On the Trusted Use of Large-Scale Personal Data. IEEE Data Eng. Bull. 35.4 (2012): 5-8.

b12. Montjoye, Y. D., Shmueli, E., Wang, S. S., and Pentland, A. S, openPDS: Protecting the Privacy of Metadata through SafeAnswers, 2014.

b13. Pentland, A.S., Social Physics: How Social Networks Can Make Us Smarter, Penguin Books, 2014: 225-233.

b14. Hardjono T, Shrier D, Pentland A. TRUST:: DATA: A New Framework for Identity and Data Sharing. Visionary Future LLC. 2016.

Recent News

News: The Human Strategy
Tue, 10/31/2017