Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines

David J. Atkinson
Institute for Human and Machine Cognition, 15 SE Osceola Avenue, Ocala, FL 34471
datkinson@ihmc.us

Abstract

The central thesis of this paper is that the technology of intelligent, autonomous machines gives rise to novel fault modes that are not seen in other types of automation. As a consequence, autonomous systems provide new vectors for cyber-attack with the potential consequence of subversion, degraded behavior or outright failure of the autonomous system. While we can only pursue the analogy so far, maladaptive behavior and the other symptoms of these fault modes in some cases may resemble those found in humans. The term "psychopathology" is applied to fault modes of the human mind, but as yet we have no equivalent area of study for intelligent, autonomous machines. This area requires further study in order to document and explain the symptoms of unique faults in intelligent systems, whether they occur in nominal conditions or as a result of an outside, purposeful attack. By analyzing algorithms, architectures and what can go wrong with autonomous machines, we may a) gain insight into mechanisms of intelligence; b) learn how to design out, work around or otherwise mitigate these new failure modes; c) identify potential new cyber-security risks; d) increase the trustworthiness of machine intelligence. Vigilance and attention management mechanisms are identified as specific areas of risk.

Introduction

Psychopathology is the study of mental illness, mental distress, and abnormal or maladaptive behavior. It is the study of *fault modes* of the human mind. As yet, we have no equivalent area of study for intelligent, autonomous machines. Software engineering techniques for reliable systems are applicable (as they are to all complex software artifacts), but insufficient. The topic of this paper is the proposition that the technology of intelligent, autonomous systems gives rise to novel fault modes that are not seen in other types of automation. These fault modes arise from the *nature of the algorithms* and how they perform in real-world situations (including human interaction) with uncertain data. As a consequence, autonomous systems may provide new vectors for cyber-attack that could lead to subversion, degraded behavior or outright system failure.

This paper arose from a bit of fun the author was having by examining examples of "robots run amok" in popular literature and media. HAL 9000, of the movie "2001: A Space Odyssey" is a canonical example. These cases are often described in anthropomorphic terms related to human psychopathology, and this became the genesis of the idea for a psychopathology of intelligent machines. Although the analogy will stretch only so far, the search for intelligent machine near-equivalents of certain human mental disorders has already yielded a few insights that are described herein. The over-riding question is whether something like the behavior of these fictional malevolent machines could actually occur. In many cases, the answer is "probably not" but in a few, the answer is "probably yes." If so, can we identify plausible mechanisms that explain the nature of the amok machines' failures, given present artificial intelligence technology and what we can reasonably project on the horizon?

This possibility suggests that there are fault modes for autonomous systems that remain unexplored and their implications unknown. The purpose of this paper is to raise that question explicitly, and to do so in the context of fault modes as potential vulnerabilities to attack, exploitation and subversion.

We stand to gain certain benefits by analyzing the unique fault modes of autonomous systems. Such studies might provide insight into aspects of machine intelligence just as studies of human mental disorders have historically provided insight into the functioning of the brain. With the human mind, psychologists seek explanations for mental disorders from biological sources (relatively rare), innate biases, and faulty inference. Such failures of the human

mind are often based in experience and learned behavior, including interpersonal communication and relationships with social and group effects. These are well documented. In contrast, with autonomous systems we must seek explanations for anomalous, maladaptive behavior in hardware (probably rare), software algorithms, logic, knowledge and situational uncertainty. Also guided by the study of human mental disorders, we should look for sources of machine intelligence fault modes in experience (episodic memory) and machine learning, including human-machine interaction and other aspects of social and affective computing.

Some of these autonomous system faults may occur in the course of day-to-day nominal operations and be easily "cured." Of greater concern, it is possible that some psychopathologies of machine intelligence could be *induced* in a new form of cyber-attack, thereby creating new risks with potentially very serious consequences. We have the opportunity, now, to focus research on how to design out, work around or otherwise mitigate the failure modes we discover. It is best if this is accomplished sooner rather than later due to the potential adverse consequences. Ultimately, the real payoff for AI research and development of autonomy applications is the opportunity to increase the trustworthiness of machine intelligence. Today, this is cited as a chief obstacle to greater deployment of autonomous systems (Dahm 2010).

The sections below provide essential background and an initial analysis of the symptoms and sources of selected example fault modes of autonomous, intelligent systems. In each case, we examine these fault modes with respect to vulnerability to cyber-attack. In the conclusion section, we discuss directions for future research and parameters of the required studies.

Essential Background

The technology of autonomous systems extends beyond conventional automation and solves application problems using materially different algorithms and software system architectures. This technology is a result of multidisciplinary research primarily in the fields of artificial intelligence and robotics, but drawing on many other disciplines as well, including psychology, biology, mathematics and others. Research on autonomous systems spans multiple areas, including (but not limited to) algorithms, computing theory, computing hardware and software, system architectures, sensing and perception, learning, and the acquisition and use of large stores of highly interconnected and structured, heterogeneous information.

The key benefit realized from autonomy technology is the ability of an autonomous system to explore the possibilities for action and decide "what to do next" with little or no human involvement, and to do so in unstructured situations which may possess significant uncertainty. This process is, in practice, indeterminate in that we cannot foresee all possible relevant information (i.e., features and their relationship to one another) that could be a factor in pattern-directed decision-making.

The autonomous ability to decide on next-steps is the core of what enables many valuable applications. "What to do next" may include a wide variety of actions, such as: a step in problem solving, a change in attention, the creation or pursuit of a goal, and many other activities both internal to the operation of the system as well as actions in the real world (especially in the case of embedded or cyber-physical systems). Ill-informed efforts to "envelope" or otherwise externally constrain the behavior of autonomous systems are sacrificing the most important strength of the technology – to perform in ways we cannot *a priori* anticipate.

However, while the technology delivers new capabilities to perform work in a wide variety of under-specified and dynamic situations, it is also extremely complex to the point where conventional software systems test and evaluation methods are no longer sufficient to establish nor maintain confidence in autonomous systems. It is system *complexity*, arising from specific component technologies of autonomy (individually and collectively), that creates the prospect of new cyber-security risks.

Of special importance is *computational complexity*: a measure of the resources required by a given algorithm to reach a result. Computational complexity is measured in time (e.g., wall clock time) and space (e.g., memory storage), and there are multiple other important attributes as well. The decision by an autonomous system of "what to do next" is the result of an algorithm that can be viewed, abstractly, as maximizing a utility function. These algorithms, intrinsic to autonomous systems, are typically of very high computational complexity; that is, they may require *exponential* amounts of time and/or space.

Strict utility-based decision-making processes are recognized to be impossible in non-trivial domains (for people as well as machines). This is a result of the potentially infinite courses of action available, and the consequent inability to exhaustively analyze all of the near infinite number of possible system states; the inability to obtain and store all potentially relevant facts; and the intrinsically uncertain relationship between chosen actions and consequences when the environment has complex dynamics including other actors (Brundage 2014).

Consequently, as a rule, the process of decision-making by an autonomous system is intrinsically limited by the available information, computational resources, and the finite amount of time available to reach a conclusion. This is referred to as "bounded rationality" (Simon 1958) and

serves as a bedrock principle for research in artificial intelligence (AI). The result is that we can only hope to *approximate* optimal decision-making and behavior in an intelligent, autonomous system: "Satisficing" is acting in a way that leads to satisfactory and sufficient ("good enough") outcomes.

We conjecture that this heuristic, *algorithmic struggle for computational resources* with limited time and information is a principal source of novel fault modes that arise in autonomous, intelligent systems.

Fault Modes

What could possibly go wrong? That is the question asked by every researcher, developer, decision-maker, and user of an intelligent system. There exists the familiar panoply of software and system faults shared by all complex computational systems. Those are not our focus here. Our interest is in what *new* types of faults might exist *by virtue of the nature of the algorithms* in intelligent systems, or their application in certain circumstances, or as a result of *malicious manipulation*. Do such fault modes exist?

The purpose of this section is to stimulate thought, discussion, and ideally, to convince you that the answer is likely to be "yes." The existence of these fault modes arises directly from the limitations imposed on autonomy technology by computational complexity, as discussed in the previous section. Such faults are today typically conceptualized in terms of constraints on algorithms rather than cyber-security vulnerabilities; this paper aims to raise awareness of that gap in our understanding.

The systems test and evaluation community has recognized that something is really different about autonomous systems, specifically, the *near infinite number of potential system states* in an intelligent, autonomous system renders much of existing test and evaluation methodology insufficient (or at worse, ineffective) for producing high confidence assertions of performance and reliability (Dahm 2010). The ideas presented here ideally ought to lead to enhanced test and evaluation processes, but we leave that to be discussed elsewhere.

In the search for novel fault modes, we are guided by our (admittedly imperfect) analogy to human psychopathology and certain philosophical considerations. If the computational mechanisms of intelligence are independent of the physical medium that supports such computations, then what is true of one type of intelligent system may also be true of another type. This is implied by the philosophical formulation of machine-state functionalism (Putnam 1979) upon which much of artificial intelligence *and* cognitive science research is predicated.

The subsections below describe potential fault modes that may arise in an example set of functional areas common to many intelligent, autonomous systems. In each case, we would like to understand the symptomology of faults and the underlying causes. Only then can we investigate vulnerabilities, methods of detection, isolation and repair. Without presenting tutorial information best found elsewhere, we consider potential fault modes arising in the processes of:

1. Goals and Goal Generation
2. Inference and Reasoning
3. Planning and Execution Control
4. Knowledge and Belief
5. Learning

Goals and Goal Generation

Goals are the primal initiator of behavior in a *deliberative* autonomous system (in contrast to a reactive autonomous system, for example, one based on a subsumption architecture (Brooks 1988) which is driven more directly by sensory data; many autonomous systems are hybrids of deliberative and reactive components. In deliberative systems, a goal state may be completely specified, only partially specified, or may be in the form of a general preference or constraint model with "goodness" evaluated according to certain formulae. A wide variety of preference/constraint models exist, some applicable only to deterministic domains and others to probabilistic domains or where preferences must be explicitly elicted (Gelain et. al. 2009; Dalla Pozza 2011).

Some examples of candidate psychopathological fault modes related to goals that are shared with people, but not other non-intelligent machines, are: Disorders of Attention, Goal Conflict, Indifference, and Self-Motivated Behavior. We examine each of these in turn.

Disorders of Attention. The pursuit of goals, including goal generation, goal selection, and deliberative planning, all require allocation of system resources. In each of these functions, decisions are made about how to use computational resources. These decision-making processes can be viewed fundamentally as *attention management mechanisms* (Helgason, Nivel and Thorisson 2012).

Goal generation functions (triggered by external or internal information) are fundamentally *vigilance mechanisms* because they can divert attention. Diverting attention diverts the management of scarce system resources. In most cases, this is appropriate and exactly what the designers of intelligent systems intend (Coddington 2007; Hawes 2011).

With respect to cyber-security, however, this suggests that attacking vigilance mechanisms has the potential to divert attention and resources away from what an autonomous system "ought" to concentrating upon. Misappropriation or diversion of scarce computing

resources is a potential critical vulnerability of autonomous systems that may appear as a consequence of other types of faults.

Goal Conflict. The resolution of conflicting requirements for achieving different goals is a fundamental component of all AI planning and scheduling algorithms. There are many such planning algorithms, and equally many ways to resolve goal conflicts. It is important to remember that *the guaranteed detection of goal conflicts during the planning process is computationally intractable*. Heuristic methods are required in order to focus attention on likely sources of goal conflict (Luria 1987). These heuristics are also *attention management mechanisms*. Luria (op. cit.) provides a brief taxonomy of goal conflicts. Drawing from that taxonomy provides a good start towards identifying goal conflict-related fault modes (see Table 1 for examples). These modes are each potential vectors for cyber-attack by an adversary with the capability to artificially induce the conditions that enable a type of goal conflict.

TYPE OF CONFLICT	DESCRIPTION
Compromised Intent	Conflict between explicit goal and default policy or implicit intent.
Violated Defaults	Unverified knowledge of the values of default preconditions.
Unintended Effects	Plan used in a novel situation with un-modeled direct interactions.
Expressed Conflict	Human agent asserts that a conflict exists, with or without explanation.
Effects Cascade	Effects of plan execution result in an unrelated conflict (side effect), e.g., due to insufficient causal model fidelity, inference horizon, etc. If the effects are non-linear, a cascade is possible.

Table 1: Example Types of Goal Conflicts.

Consider just one of the many sources of goal conflicts that are known: *Compromised Intent*. This type of goal conflict occurs when achievement of a goal conflicts with default policy or intent. It may occur because (1) a causal interaction is not modeled, or; (2) an inference chain is too long to find the conflict (as in a search with a bounded horizon), or; (3) unknown, explicit or implicit priorities, or other conditions that enable the relaxation of constraints. I would be greatly surprised if a reader familiar with AI planning systems has not seen this type of conflict.

There is another reason it seems familiar. Return to our (fictional) example in the introduction, HAL 9000, of the movie "2001: A Space Odyssey." Recall that the super-secret, highest priority mission goal given to HAL is to investigate the monolith at Jupiter. This explicit goal comes into conflict, later in the mission, with the default policy of protecting the lives of the crew. This is just one of many types of potential goal conflicts that may not be detected before actual execution of a plan. Skipping over the drama of the movie, we discover that HAL chooses to resolve this goal conflict by killing the crew. The hypothetical mechanism is relatively easy to discern: a relaxation of a constraining default policy (crew safety) in order to achieve a high priority goal (investigate the monolith). The constraint relaxation is enabled by HAL's *certainty* that he can complete the secret mission without the aid of the crew (this is also a failure of *ethical reasoning*, discussed later). In humans, unresolved goal conflict is a source of significant mental distress (Mansell 2005). Similarly, resolution of goal conflict is often (but not always) an imperative in autonomous systems.

Indifference. This type of fault is a milder form of goal conflict that can result from an intelligent system concluding that (1) a human-provided goal has insufficient priority relative to other goals, or (2) the system itself does not possess the competence to achieve a goal, or (3) the goal is irrelevant. The consequence of goal conflict resolution is the human-provided goal is dropped and no action occurs to achieve the goal. In human terms, this condition is described as *apathy*.

Self-Motivated Behavior. Autonomy necessarily implies a degree of choice of actions, whether they originate from externally provided goals or goals internally generated. In the latter case, the addition of autonomic processes to a system (e.g., for health maintenance, energy management and so forth) can result in goals that conflict with on-going activities. The examples we see today, such as a robot vacuum cleaner stopping to recharge itself, are expected behaviors and not of interest. However, as intelligent robots are deployed into dangerous situations, such as urban rescue or a battlefield, their autonomic functions are likely to expand to include *self-preservation* as a default autonomic function. Consider the possibility that an undesirable machine behavior (perhaps as a result of another fault and/or subversion) results in a shutdown command. Due to a conflict with the internally generated goal of self-preservation, the directive to shutdown could be ignored in certain situation-driven conflict resolutions.

It is important to note that both *Indifference* and *Self-Motivated Behavior* may have the *appearance* of self-awareness. Yet, the information and goal conflict resolution processes are localized within the intelligent system. The appearance of self-awareness is an *emergent psychological effect* (Lewis et. al. 2011); actual self-awareness is not required.

Inference and Reasoning

There are several important sources of potential faults in the area of inference and reasoning that require further

study. Some are familiar to many of us from long coding and test sessions with intelligent systems, others are speculative possibilities that may arise in future systems.

Invalid Logic. Often termed "fallacies of inference", there are many forms of invalid logic that humans demonstrate. As yet, intelligent systems only suffer from a few. One of these cases is when "true" data results in a false answer as a result of a failure of inductive reasoning; for example, when an intelligent system is near the edges of its competence. As a consequence, insufficient previous experience (e.g., manifested as an incorrect probability distribution) results in over-weighted confidence for derived conclusions. This leads to the possibility that new data becomes marginalized or discarded rather than serving in a corrective function. This is an example of the classic "over-generalization" problem in machine learning, where important features that discriminate situations are ignored.

The Fallacy Fallacy. This fault mode is complementary to *Invalid Logic*. Knowledge bases are inherently incomplete, likely to contain errors, and subject to many other limitations. One potential consequence is that a conclusion is dismissed because the logic used to derive the conclusion is faulty or incomplete, i.e., there is no inference chain to the conclusion that can be constructed from the knowledge and data given (or as a result of a bounded search horizon, as discussed earlier). If the argument contains a fallacy, i.e., invalid logic, then *it is the argument that must be dismissed*. The failure to construct an inference chain does not prove that the *conclusion* is incorrect, only that it cannot be proved with what is known. The conclusion may in fact be correct. Few, if any, extent intelligent systems respect this distinction; it is a defect of reasoning that unfortunately shared by many people as well.

Solipsism. One of the dangers of the AI craft of applied epistemology arises from the quest to manage uncertainty. This has the potential to result in a sort of *logical minimalism* where sense data is subject to extreme skepticism and as a result, internally derived inferences may accrue more confidence than those based on empirical observations. In a sense, this is the robot equivalent of the human psychopathological condition of *detachment from reality*. The danger arises when solipsism undercuts externally imposed policy-guided constraints on behavior by authority.

Planning and Execution Control

There are a great many faults that can arise during the planning and execution control processes, including many of those related to goals as we have discussed above. Planning is essentially a search problem with surprising complexity that often requires exponential computation, i.e., is NP-hard (Chapman 1987; Hendler et. al. 1990) One of the most important potential fault modes of planning and execution control has only been recognized in the past few years: failures of *ethical behavior* (Arkin 2012; Bringsjord and Clark 2012).

Ethical reasoning may fail due to bounded rationality. Depending on the circumstances, knowledge and analysis of the situation and actors may not be sufficient to reason about duty to ethical concerns. It is also true that creating an *ethical code* that is *complete*, *unambiguous* and can be *applied correctly* in every situation is notoriously difficult (Bringsjord 2006). Many possible algorithms to remedy this have been discussed (and fewer implemented), such as "ethics governors" (an execution monitoring system with veto power; essentially equivalent to the proverbial "restraining bolt"). Other theorists suggest that moral behavior will arise not from externally imposed constraints, but only from internally generated self-regulation of behavior based on the utilitarian concerns of interacting with humans in a social world.

A discussion of ethical behavior by machines is not complete without a consideration of *deception*, defined for our purposes here as a "false communication that tends to benefit the communicator" (Bond and Robinson 1988). For reasons of space, a complete review is not possible here. With respect to our concerns regarding fault modes and cyber-security, it is important to note that deception by an autonomous, intelligent system can arise naturally as an adaptive response to certain situational conditions (Floreano 2007; Mitri 2009), as a strategic choice, e.g., in warfare (Wagner and Arkin 2009, 2011), or as a relatively innocuous aspect of human-robot social interaction (Pearce et. al. 2014). This raises the question *of how to tell the difference between a mistake* (due to a failure or limitation) *and an outright lie* by an intelligent system.

While there is much attention to policy-constrained behavior (Uszok et. al. 2008), The fact remains that today we cannot *guarantee* that the behavior of a sufficiently autonomous intelligent system will necessarily conform to explicitly stated policies, including ethical rules. The consequences might be relatively minor or they might be as major as the HAL 9000 goal conflict resolution example discussed earlier.

Emergent Behavior. As the technology of multi-agent systems has matured, the phenomenon known as emergent behavior has been observed, i.e., behavior that is not attributed to any individual agent, but is a global outcome of agent coordination (Li et. al. 2006). Emergent behavior may or may not represent a fault condition. The flocking behavior of birds is emergent and represents an important positive survival trait. On the other hand, stop and go traffic and traffic jams are emergent behaviors that enormously degrade the performance of traffic systems.

As yet, no generalizable methods exist for predicting emergent behavior in multi-agent systems, or their "goodness", in part because the task is computationally

intractable even for very simple agents with restricted behavioral repertoire and restricted inter-agent communication topology. Emergent behavior cannot be predicted by analysis at any level other than the system as a whole. The best that can be done is to measure certain trends in system-wide behavior that may lead to predictability (Gatti et. al. 2008; Pais 2012).

A fault mode worthy of study is the possibility that *an agent in a multi-agent system is able to assert its behavior on other agents in a way that triggers emergent effects* (Lewis et. al. 2011). Two simple examples, similar in nature, are crowd behavior in humans and insect swarming. To the extent that an agent suffers some other fault, or is suborned, it may trigger undesirable emergent behaviors in the system as a whole. Despite the rush to implement multi-agent systems for important and critical applications in health, finance, transportation, defense and other domains, we simply do not yet have an understanding of fault modes that are likely to occur due to emergent behavior.

Learning, Knowledge and Belief

This category of potential fault modes is quite broad and truly deserves more attention than this short paper can afford. Nevertheless, it is important to highlight a few fault modes that may be quite common to intelligent, autonomous systems. These all arise from the autonomous processes involved in creating, maintaining, and adapting what an intelligent system believes to be true.

The most glaring example of this type of fault mode is faulty or absent truth maintenance, i.e., the ability to retract assertions previously thought to be true which are now rendered invalid by new information (defeasibility). Formally, this is a property of first order logical "monotonic" systems. The use of monotonic inference is not in itself a fault. If previous inferred assertions do not play a role in future reasoning, they are effectively discarded if not explicitly falsified when contradictory information is obtained. For example, a credit card fraud detection system might depend exclusively on salient features in a single case of use of the card. The fact that a previous use of the card was valid does not automatically validate a new use of the card. First order logic is common in many applications.

However, intelligent systems that build models of the world, actors, situations, and so forth via machine learning must use *non-monotonic reasoning* (second order or higher logics) to achieve defeasible inference. Given the uncertainty inherent in a dynamic and uncertain world, defeasibility can be a difficult process because it requires weighing the evidentiary force of new data against previously derived probative assertions. In a sense, skepticism must balance a rush to learn or "correct" previous beliefs.

This is where computational argumentation and its contribution to persuasive technology may have an important role. While the topics are strongly related to formal logic and mathematical proof, they transcend it in several ways. Most important to this discussion is *the explicit inclusion of dialog in the process of argumentation*, often in the context of creating "explanations" as to why certain conclusions have been reached, as in intelligent decision-support systems (Bench-Capon et. al. 1991,2007a). In this context, argumentative dialog is an *exchange* of ideas using rhetorical methods of persuasion that include social methods as well as mathematical logic.

Justification of belief though argumentative dialog opens the door to fallacious reasoning as a method of persuasion. "Appeal to Authority" (*argumentum ad verecundiam*), while regarded as fallacious in theories of debate, cannot be ruled invalid simply by noticing it in dialog – it requires a further exchange of ideas. In the absence of effective counter argument, by either human or machine participant, *fallacious reasoning may be highly influential* as a result of "practical reasoning", i.e., an assertion is correct within the perspective of one of the agents involved (Bench-Capon and Dunne 2007b). Humans are particularly vulnerable to deceptive cognitive illusions that result from certain argumentation strategies and practical reasoning. The computational methods for exploiting this weakness are actively being explored (Clark and Bringsjord 2008).

The cyber-security concern is that practical reasoning to justify belief in the presence of uncertainty opens the door to the possibility that an adversary could, though the argumentative dialog process, *undermine an intelligent system's beliefs*. This would be an even greater risk in the context of supervised learning with training data. Supervised machine learning is already known to be subject to a number of systematic biases, including for example, order bias, recency bias, and frequency bias. Errors in causal attribution can easily result from these biases.

A second, related cyber-security concern is the role practical reasoning could play in goal generation and planning. By undermining (or cunningly shaping) an autonomous, intelligent system's beliefs, all of the goal-related fault modes discussed earlier could be induced.

Conclusions

Inherent in the concept of autonomy in intelligent systems is the ability to make choices about what to do and how to do it. These are fundamentally mechanisms for managing attention and vigilance. In this paper, we have examined some of the components of intelligent systems that support autonomy and discussed a selection of potential fault

modes. Some of these fault modes require a degree of meta-cognition that, while not yet realized in autonomous systems, is an active area of research.

It is possible that some or all of these fault modes can be induced, and as a consequence, there now exist new and unique cyber-security concerns surrounding autonomous systems that must be explored. It is therefore incumbent on the AI research community to establish a theoretical and empirically substantiated foundation for cyber-security issues related to autonomy, with special attention to gaps in current knowledge. Future studies of cyber-attack vulnerabilities, per fault modes that are related to autonomy, should explore the following:

1. **Fault Modes**: Are there new types of fault modes that can be exploited? Which fault modes are possible to induce, and in what manner and circumstance?
2. **Detection**: How can we detect that an intelligent, autonomous system has been/is being subverted?
3. **Isolation**: In the context of autonomous system faults and possible subversion, what do the traditional system concepts of *fail safe* and *fail operational* mean?
4. **Resilience and Repair**: What are the proximal causes of the observable symptoms of autonomous system fault modes and how can these be mitigated?
5. **Consequences of Vulnerabilities**: What are the consequences of deception by an autonomous, intelligent system (whether it has been subverted or not)? What is the impact of different types of fault modes on human reliance, trust, and performance of human-machine systems?

The inspiration for this paper was the question of whether fictional dramatic accounts of computers/robots "run amok," often described in anthropomorphic terms, have the potential to actually occur either with existing technology or technology that can be reasonably foreseen on the horizon. Some, but not all, of these faults and vulnerabilities have useful analogies to psychopathologies of the human mind. The development of a theory of "psychopathology of intelligent machines" has the potential to provide insight into aspects of computational intelligence just as studies of human mental disorders provide insight into the functioning of the brain. The methodology that remains to be developed will guide us towards computational approaches that design out, work around or otherwise mitigate these failure modes and potential cyber-security risks. Ultimately, the real payoff is the opportunity to increase the trustworthiness of machine intelligence; in the absence of justifiable trust, the full potential of autonomy technology will not be realized.

Acknowledgements

The author would like to thank Micah Clark and his other colleagues at IHMC for their fruitful discussions of the topics discussed in this paper. This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-12-1-0097.

References

Arkin, R.C. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception. *PROC IEEE.* 100(3):571-589.

Bond, C. F., & Robinson, M. 1988. The evolution of deception. *Journal of Nonverbal Behavior.* 12(4):295-307.

Bench-Capon, T.J.M., Dunne, P.E. and Leng, P.H. 1991. Interacting with knowledge-based systems through dialogue games. In *Proc. 11th Annual Conf. Expert Systems and their Applications*. 123–130.

Bench-Capon, T.J.M., Doutre, S. and Dunne, P.E. 2007a. Audiences in argumentation frameworks. *Artificial Intelligence.* 171:42–71.

Bench-Capon, T.J.M. and Dunne, P.E. 2007b. Argumentation in artificial intelligence. *Artificial Intelligence.* 171:619-641.

Bringsjord, S., Arkoudas, K. and Bello, P. 2006. Towards a General Logicist Methodology for Engineering Ethically Correct Robots. *Intelligent Systems, IEEE.* 21(4):38-44.

Bringsjord, S., and Clark, M. 2012. Red-Pill Robots Only, Please. *IEEE Trans. Affect Comput.* 3(4):394–397.

Brooks, R.A. 1988. A robust layered control system for a mobile robot. *IEEE J ROBOT AUTOM.* (2):14–23.

Brundage, M. 2014. Limitations and Risks of Machine Ethics. Technical Report from Consortium for Science, Policy, and Outcomes. Tempe, AZ: Arizona State University.

Chapman, D., and Agre, P. 1987. Abstract Reasoning as Emergent from Concrete Activity. In *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop.* Eds. M. Georgeff and A. Lansky. San Mateo, CA: Morgan Kaufmann.

Clark, M. and Bringsjord, S. 2008. Persuasion Technology Through Mechanical Sophistry. *Communication and Social Intelligence.* 51-54.

Coddington, A. 2007. Motivations as a Meta-level Component for Constraining Goal Generation. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems.* 850-852.

Dahm, W. 2010. Technology Horizons: A Vision for Air Force Science & Technology During 2010–2030. *Technical Report AF/ST-TR-10-01-PR*. Washington, DC: United States Air Force, Office of Chief Scientist (AF/ST).

Dalla Pozza, G., Rossi, F. and Venable, K.B. 2011. Multi-agent Soft Constraint Aggregation – Sequential approach. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence.* Morgan Kaufmann. 1.

Floreano, D., Mitri, S., Magnenat, S., & Keller, L. 2007. Evolutionary Conditions for the Emergence of Communication in Robots. *Current Biology.* 17(6):514-519.

Gatti, M.A., Lucena, C.J., Alencar, P. and Cowan, D. 2008. Self-Organization and Emergent Behavior in Multi-Agents Systems: A

Bio-inspired Method and Representation Model. In *Monografias em Ciência da Computação.* 19(8). ISSN: 0103-9741.

Gelain M., Pini, M.S., Rossi, F., Venable, K.B. and Walsh, T. 2007. Elicitation Strategies for Soft Constraint Problems with Missing Preferences: Properties, Algorithms and Experimental Studies. *Artificial Intelligence.* Elsevier. 174(3-4):270-294.

Handler, J., Tate, A. and Drummond, M. 1990. AI Planning: Systems and Techniques. *AI Magazine* 11(2):61-77.

Hawes, N. 2011. A survey of motivation frameworks for intelligent systems. *Artificial Intelligence.* 175:1020-1036.

Helgason, H.P., Nivel, E. and Thorisson, K.R. 2012. On Attention Mechanisms for AGI Architectures: A Design Proposal. *Artificial General Intelligence*. Lecture Notes in Computer Science. Springer. 7716:89-98.

Lewis, P. Chandra, A., Parsons, S., Robinson, E. et. al. 2011. A Survey of Self-Awareness and Its Application in Computing Systems. In *Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops*. IEEE Comp. Soc. 102-107.

Li, Z., Sim, C.H., and Low, M.Y.H. 2006. A Survey of Emergent Behavior and Its Impacts in Agent-Based Systems. In *Industrial Informatics, Proceedings of the 2006 IEEE International Conference on Industrial Economics.* 1:1295-1300.

Luria, M. 1987. Goal Conflict Concerns. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*. Milan, Italy: Morgan Kaufmann. 2:1025-1031.

Mansel, W. 2005. Control theory and psychopathology: an integrative approach. *PSYCHOL PSYCHOLTHER.* 78(2):141-178.

Mitri, S, Floreano, D. and Keller L. 2009. The evolution of information suppression in communicating robots with conflicting interests. *Proceedings of the National Academy of Sciencies.* 106(37):15786-15790.

Pais, D. 2012. Emergent Collective Behavior in Multi-Agent Systems: An Evolutionary Perspective. Ph.D Dissertation. Princeton University.

Pearce, C., Meadows, B., Langley, P and Burley, M. 2014. Social Planning: Achieving Goals by Altering Others' Mental States. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.* 1:402-408.

Putnam, H. 1979. *Philosophical Papers: Volume 2, Mind, Language and Reality*. Boston, MA: Cambridge University Press.

Simon, H. 1958. Models of Man. *Journal of the American Statistical Association* 53(282):600-603. Taylor & Francis, Ltd.

Uszok A., Bradshaw, J., Lott, J. et. al. 2008. New Developments in Ontology-Based Policy Management: Increasing the Practicality and Comprehensiveness of KAoS. In *Policies for Distributed Systems and Networks (POLICY 2008), Proceedings of the IEEE Workshop on.* IEEE Press. 145-152.

Wagner, A.R. and Arkin, R.C. 2009. Deception: Recognizing when a Robot Should Deceive. *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*. IEEE Press. 46–54.

Wagner, A.R. and Arkin, R.C. 2011. Acting Deceptively: Providing Robots with the Capacity for Deception. *Int J Soc Robotics.* Springer. 1-22.