FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

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

David J. Atkinson

Senior Research Scientist

Institute for Human and Machine Cognition

AAAI Spring Symposium Series:
Foundations of Autonomy and Its (Cyber) Threats
Palo Alto, CA March 2015

Topics

- Definitions
- Hypothesis
- Essential Background
- Fault Modes of Minds
  - Analysis, Methodology
- Examples
- What is Needed
- Next Steps

Definitions

Autonomy: The ability of an intelligent agent to explore the possibilities for action and decide "what to do next" with little or no human involvement, even in unstructured situations where there exists significant uncertainty

Agent: A self-activating, self-sufficient and persistent computation that is capable of modifying the manner in which it achieves objectives and fulfills its purpose

Autonomous Agents
are becoming Cognitive Entities

- They have goals that we give them and goals of their own
- The have beliefs about themselves, the world, and others
- They sense and interpret the world, reason in many different ways about their beliefs, and act purposefully to achieve goals
- They learn and adapt
- They interact with humans, perform significant actions in the world, communicate in a variety of ways, and are beginning to act socially

Hypothesis

The technology of intelligent, autonomous agents gives rise to novel fault modes not seen in ordinary software or automation

- Inspiration: Philosophy of Functionalism
  - Any creature with a mind can be regarded as a Turing machine (an idealized finite state digital computer)
  - Mental states are best regarded as computational states
- Implication
  - ***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

Psychopathology: Fault Modes of Minds

- Traditionally, the study of disorders of the human mind
  - Biological sources (rare)
  - Psychological sources (most common)
    - Cognitive and Affective Processes
    - Experience, learned behavior, communication, relationships
  - Symptoms: biases, faulty inference, and an array of mental and behavioral problems

- No equivalent area of study for disorders of artificial minds
  - Hardware sources (probably rare)
  - Algorithm sources (probably more common)
  - Knowledge, learning, social interaction and uncertainty (probably more common)

Symptoms: Biases, Faulty Inference, and an Array of Mental and Behavioral Anomalies

Potential Consequences

- Unanticipated "Maladaptive" Behaviors
  - Emergent disorders
- Novel Vulnerabilities to Cyber-Attack
  - Subversion
  - Induced Degradation or Failure

Methodology

- Functionalism suggests an exploratory methodology; be guided by what we know of human minds
- Forecast and search for autonomy fault modes whose symptomology is similar to human psychopathologies
- Leverage current understanding of human psychopathologies to hypothesize mechanisms for autonomy fault modes

Examples of Autonomy Functions Needing Investigation

- Goals and Goal Generation
  - Disorders of Attention
  - Goal Conflict
  - Indifference
  - Self-Motivated Behavior
- Inference and Reasoning
  - Invalid Logic (Fallacy)
  - The Fallacy Fallacy
  - Solipsism
- Planning and Execution Control
  - Ethical Behavior
Emergent Behavior
- Learning, Knowledge, Belief
  - Failures of Truth Maintenance
  - Practical Reasoning Undermined by Argumentation
  - Biased Experience
  - Biased Training Data

Example #1

Goals and Goal Generation

- A core process in all deliberative systems
  - Initiates both internal (pattern-directed computation) and external behaviors (actions in real world)
- Candidate fault modes in Goal Conflict
  - Compromised intent
  - Violated defaults
  - Unintended effects
  - Expressed conflict
  - Effects cascade
- Guaranteed detection during planning process is **intractable**
  - **Heuristic "Attention Management Mechanisms" focus search**

Example #1

Compromised Intent

- Achievement of a goal conflicts with a default policy or intent
  - A causal interaction is not modeled
  - An inference chain is too long to find the conflict (e.g., due to bounded horizon)
  - Unknown, explicit or implicit priorities, or other conditions enable inappropriate relaxation of constraints.
- Why does this seem familiar?

HAL 9000 from *2001: A Space Odyssey*

- A high priority goal
- Conflicts with default policy
- A belief enables alternative method of achieve goal
- Default policy constraint is relaxed and the conflict resolved

Example #2

Inference and Reasoning

- Process to determine the truth value of assertions and to reveal the logical implications that follow
  - Related to Goals and Goal Generation

- Candidate fault modes in inference and reasoning
  - Invalid Logic (Fallacies of Inference)
  - The Fallacy Fallacy
  - Solipsism

- Management of uncertainty

Example #2

The Fallacy Fallacy

- A conclusion is dismissed because the logic or knowledge used to derive it is faulty or incomplete
  - Knowledge bases are buggy and incomplete
  - Bounded search horizons may limit length of inference chains
  - The inference chain included fallacies
  - May also be related to a failure of truth maintenance (defeasibility)

- If the argument is incomplete or faulty, then it is the **argument** that must be dismissed, not the conclusion.
  - Failure to construct and inference chain demonstrates only that the conclusion cannot be proved with what is known
  - **The conclusion may in fact be correct**

A defect in reasoning that is unfortunately common in both people and intelligent machines

Root Causes?

Bounded Rationality: Decision-making is intrinsically limited by the available *information*, computational *resources*, and the finite amount of *time* available to reach a conclusion

- "What to do next" ~= Maximizing a Utility Function
  - Strict utility-based decision processes are impossible in non-trivial domains
  - **Indeterminate**: We cannot foresee all possible relevant information and relationships that could be a factor in pattern-directed decision-making
- The Problem
  - Potential **infinite** courses of action available
  - Impossible to analyze the near-infinite number of potential system states
  - Impossible to obtain and store all potentially relevant facts
  - Uncertain relationship between actions and consequences

Conjecture: The heuristic struggle for computational resources is a principal source of novel fault modes

Cyber-Security Vulnerabilities

- Attacks on **Attention Management Mechanisms**
  - Goal stacks (and equivalent)
  - Goal prioritization and selection
  - Priming values for search heuristics

- Attacks on **Vigilance Mechanisms**
  - Goal generation
  - Uncertainty thresholds

- Deceptive Manipulation of **Policies**
  - We cannot guarantee that the behavior of an autonomous intelligent system will conform to stated policies, including ethical rules

- …Social Engineering?
  - Reverse persuasive technology to subvert machine learning

Recommendation

- The AI community should establish a theoretical and empirically substantiated foundation for cyber-security issues related to autonomy

- Studies should explore:
  - **Fault Modes**: What types can be induced, and in what circumstances:
  - **Detection**: Can we tell if a system has been subverted?
  - **Isolation**: What does "fail safe" and "fail operational" mean in this context?
  - **Resilience and Repair**: What are the proximal causes and can they be mitigated?
  - **Consequences**: Deception? Impact on human reliance, trust and performance?

Additional Benefits

- Potential insight into aspects of artificial intelligence just as studies of human mental disorders have taught us about the functioning of the human mind.

Thank You!

David J. Atkinson

datkinson@ihmc.us