First International Workshop on Artificial Intelligence Safety Engineering (WAISE), 2018

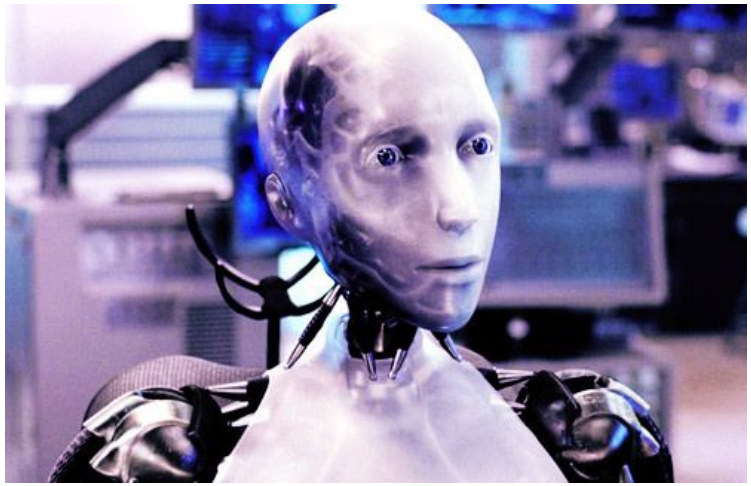
A Psychopathological Approach to Safety Engineering in AI and AGI

Vahid Behzadan, Arslan Munir
Kansas State University
and Roman Yampolskiy
University of Louisville

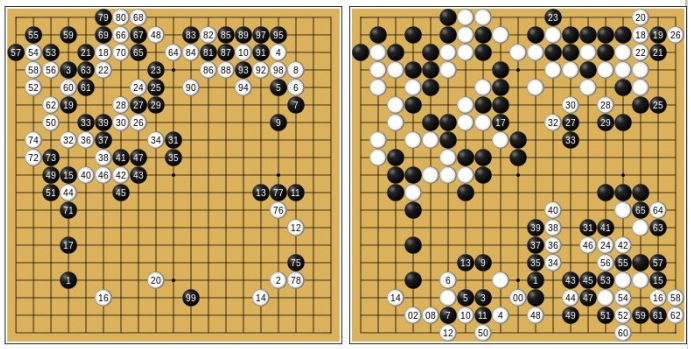
Email: behzadan@ksu.edu
http://www.vbehzadan.com

Outline

- Motivation
- What is Psychopathology?
- Psychopathology and AI Safety
- Directions of Research
- Conclusion



State of AI Safety


First 99 moves (96 at 10)

Moves 100-165.

- Complexity of dynamics
  - e.g., deep RL
  - AI/AGI ➡ complex systems
- Current trends in AI Safety
  - Major focus on safe design
  - Formal analysis (e.g., controllability, reachability)
    - Already too difficult
  - Technical abstraction (e.g., training algorithms, model architecture, etc.)
    - Already too complex
- Need for higher-level abstractions

Natural Inspiration

- Human Cognition
  - General purpose intelligence
  - Adaptive to dynamic environments
  - Prone to psychological disorders:

  Self-reconfigurations in cognition and behavior that are deleterious to the core and long-term objectives of self or the social ecosystem
- AI/AGI safety issues follow the same definition
  - Unintended [emergent] behavior that is harmful to the core and long-term objectives of designer or the society.

Psychopathology

- Scientific study of mental disorders, causes, and treatments
- Mental disorder: “behavioral pattern associated with distress, disability, increased risk of death, or significant loss of autonomy” (APA 2013)
- Identified based on 4 Ds:
  - Deviance of behavior from norm
  - Distress of the individual
  - Dysfunctions impairing ability to perform designated functions
  - Danger to self or society

Psychopathology (2)

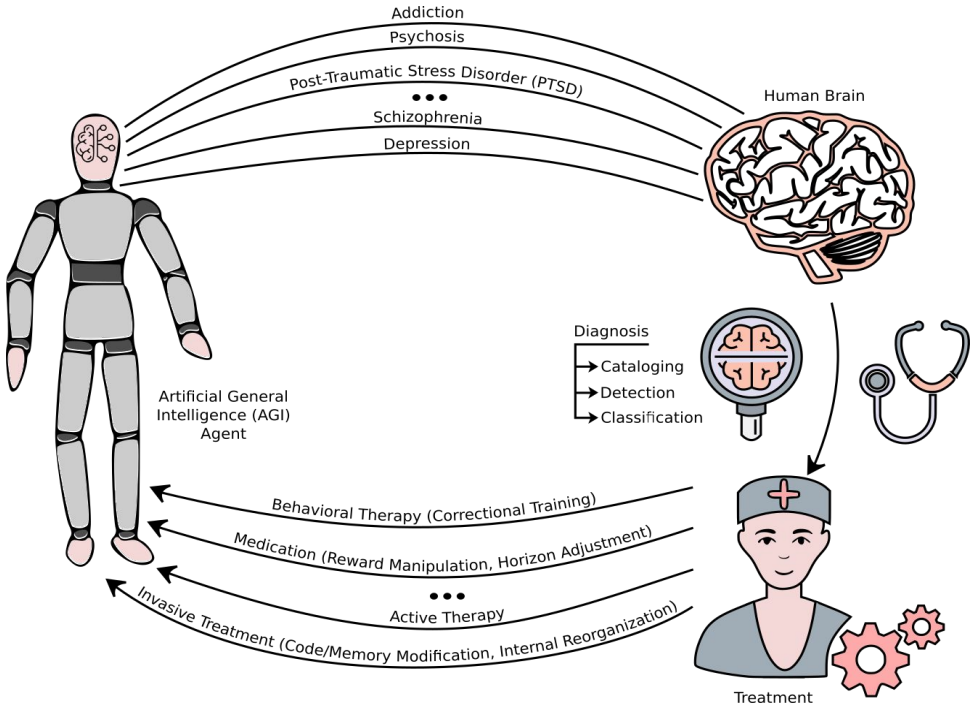
- Causes: genetic, developmental, social trauma, biological
- Diagnosis: APA's Diagnostic and Statistical Manual of Mental Disorders (DSM)
- Treatments
  - Psychotherapy (e.g., Cognitive Behavioral Therapy - CBT)
  - Medication Therapy
  - Hybrid

Psychopathology of AI

- Failures in AI safety as psychological disorders
  - Reward hacking ⟶ harmful/addictive behavior (Yampolskiy 2014)
  - Unsafe Exploration ⟶ delusional behavior (Yampolskiy 2014)
  - Negative Side Effects ⟶ psychopathy (MIT's Norman), Depression (Ashrafian 2015)
  - Robustness to Distributional Change ⟶ depression (Ashrafian 2015), PTSD (Ashrafian 2015)
  - etc...

Research Framework

- Components of AI Psychopathology:
  - Modeling and Verification Tools
  - Classification and Diagnosis of Disorders
  - Treatment



Modeling and Validation

- Mathematical modeling of analogues between human psychopathology and AI safety
  - e.g., RL-based modeling of schizophrenia (Montague et al., 2004)
- Experimental Frameworks
  - Simulation platforms
  - Benchmark experiments
  - Development/Adoption of suitable metrics

Classification and Diagnosis

- Diagnosis
  - Detection of anomalous behaviors
    - Tripwires and Honeypots (FLI 2017)
    - Statistical deviations in behavior
    - Indicators of misbehavior
  - Classify type of anomalous behavior
    - Compilation of representative and experimentally verified disorders,
    - Derivation of criteria and indicators for disorders (i.e., DSM for AI)
    - Data collection techniques: direct interaction, non-invasive analysis of internal data, invocation of internal debug modes, etc.

Treatment

- Decommission/Reset isn't always the optimal solution
  - Unique skills
  - Human-Agent affection
  - Re-training cost
- Approaches:
  - Correctional Training (parallel to behavioral therapy)
    - e.g., simulated environment to allow more extensive exploration
  - Artificial Reward Induction (parallel to medication therapy)
    - e.g., artificial induction of negative reward when undesired behavior is shown

Conclusion

- Need for higher-level abstractions for AI/AGI safety
- Psychopathology provides many practical analogues
- AI safety can build on current knowledge in psychopathology
- Research in this area can benefit both fields of AI safety and psychopathology

Questions?