A psychopathological approach to safety engineering in AI and AGI

Vahid Behzadan
Kansas State University

Arslan Munir
Kansas State University

Roman V. Yampolskiy
University of Louisville, roman.yampolskiy@louisville.edu

Original Publication Information

Behzadan V., Munir A., Yampolskiy R.V. (2018) A Psychopathological Approach to Safety Engineering in AI and AGI. In: Gallina B., Skavhaug A., Schoitsch E., Bitsch F. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2018. Lecture Notes in Computer Science, vol 11094. Springer, Cham. https://doi.org/10.1007/978-3-319-99229-7_46

ThinkIR Citation

Behzadan, Vahid; Munir, Arslan; and Yampolskiy, Roman V., "A psychopathological approach to safety engineering in AI and AGI" (2018). *Faculty Scholarship*. 572.
https://ir.library.louisville.edu/faculty/572

A Psychopathological Approach to Safety Engineering in AI and AGI

Vahid Behzadan[1], Arslan Munir[1], and Roman V. Yampolskiy[2]

[1] Kansas State University, Manhattan, KS 66506, USA
{behzadan,amunir}@ksu.edu
http://blogs.k-state.edu/aisecurityresearch/
[2] University of Louisville, Louisville, KY 40292, USA
roman.yampolskiy@louisville.edu

Abstract. The complexity of dynamics in AI techniques is already approaching that of complex adaptive systems, thus curtailing the feasibility of formal controllability and reachability analysis in the context of AI safety. It follows that the envisioned instances of Artificial General Intelligence (AGI) will also suffer from challenges of complexity. To tackle such issues, we propose the modeling of deleterious behaviors in AI and AGI as psychological disorders, thereby enabling the employment of psychopathological approaches to analysis and control of misbehaviors. Accordingly, we present a discussion on the feasibility of the psychopathological approaches to AI safety, and propose general directions for research on modeling, diagnosis, and treatment of psychological disorders in AGI.

Keywords: AI Safety · Psychopathology · Mental Disorder · Diagnosis · Treatment · Artificial General Intelligence.

1 Introduction

While the adaptive mechanisms of human cognition provide the means for unique skills in adjusting to dynamic environments, they are also prone to psychological disorders, broadly defined as self-reconfigurations in cognition and behavior that are deleterious to the core and long-term objectives of self or the social ecosystem [2]. Extrapolating from this phenomenon, it is not hard to conclude that instances of Artificial General Intelligence (AGI), which aim for similar cognitive functions, may also be prone to such disorders. For instance, certain objective functions and environmental conditions may lead a Reinforcement Learning (RL) agent to develop addictive behavior through repetitive gains of high rewards from policies that contradict the long-term objectives of the agent [15]. Other instances of such emergent disorders include post-traumatic behavior, depression, and psychosis [1]. It is further hypothesized that behavioral disorders may emerge as higher-order consequences of unsafe inverse RL and apprenticeship learning, by adopting manifested disorders or triggering harmful cognitive traits [16].

Current research in AI safety is generally focused on safety-aware design and mitigation techniques [11], but the expanding complexity of AI and in particular AGI will render such analysis as difficult as those of biological intelligence and the corresponding disorders. To tackle such difficulties in human intelligence, the causes and dynamics of misbehaviors are studied at various levels of abstraction, ranging from neuroscience and cognitive science to psychology, psychiatry, sociology, and criminology. Inspired by the advantages of such diverse vantage points, we propose that studying the complex dynamics and mechanisms of failure in AI safety can greatly benefit from abstractions that parallel those of biological intelligence. Considering the practical aims of diagnosing and correcting misbehaviors in AGI, we believe that adopting the abstraction of psychopathology provides tractable settings that also benefit from cross-domain bodies of knowledge. Furthermore, while this approach may seem to be of lower relevance at present, we argue that the advent of deep RL, along with advances in hierarchical and transfer learning may have already laid the grounds for emergence of such disorders in AI.

The goal of this paper is to provide a technical discussion and the motivation for research on the psychopathology of AI and AGI. The remainder of this paper is organized as follows: Section 2 presents a broad overview of psychopathology. Section 3 provides a discussion on the relevance of psychology to AI, followed by establishment of parallelisms between AI safety and psychopathology. In Section 4, high-level areas of research are identified and detailed. Finally, Section 5 concludes the paper with remarks on broader impacts of this research.

2 What is Psychopathology?

Psychopathology refers to the scientific study of mental disorders, their causes, and corresponding treatments [4]. Within this context, we adhere to American Psychiatric Association (APA)'s definition of mental disorder [2] as "a psychological syndrome or pattern which is associated with distress, disability, increased risk of death, or significant loss of autonomy" (i.e., pursuit of objectives). In psychopathology, disorders are commonly identified based on four metrics of abnormality, known as the four Ds [6]: Deviance of behaviors and emotions from the norm, Distress of the individual caused by suffering from a disorder, Dysfunctions that impair the individuals ability to perform designated or normal functions, and the Danger of individual to self or the society.

Causes of mental disorders in humans include mixtures of those inherited through *genetics* (e.g., neuroticism), *developmental influences* caused by parental mistreatment, social influences (e.g., as abuse, bullying), and traumatic events, and *biological influences* such as traumatic brain injury and infections [2].

Various models have been developed to capture the dynamics of mental disorders and their emergence. For instance, biological psychiatry, or the *medical model* [9], is one that explains the causes of disorders based on changes in neurological circuitry. The *social model*, on the other hand, analyzes the causes of mental disorders based on social and environmental interactions [9]. Currently, it is widely believed that understanding psychological disorders requires the comprehensive consideration of both biological and social factors, and hence the *biopsychosocial models* are generally adopted to study such phenomena. These models broadly categorize mental disorders as either cognitive or behavioral. Cognitive disorders are those caused by abnormal functioning of the underlying cognitive mechanisms, and behavioral disorders are those that are learned through developmental, environmental, and social interactions [9].

Diagnosis of mental disorders is generally based on an assessment of symptoms, signs, and impairments that constitute various types of disorders. A comprehensive framework for such assessments is that of the Diagnostic and Statistical Manual of Mental Disorders (DSM) [2], published by the American Psychiatric Association (APA). This manual provides a common language and standard criteria for the classification of mental disorders. Furthermore, recent advances in machine learning have given rise to various software and algorithmic tools to facilitate enhanced accuracy in classification and diagnosis of mental disorders [8].

Treatment of mental disorders is commonly via one or a hybrid of two approaches. One is *Psychotherapy*, which is a form of interpersonal intervention via a range of psychological techniques. For instance, Cognitive Behavioral Therapy (CBT) is employed to modify the patterns of thought and behavior associated with a particular disorder. *Medication therapy* is the other approach, which targets the physiological components of disorders. For instance, antipsychotics commonly work by blocking D2 Dopamine receptors, thus controlling the chemical reward mechanism of the brain [13].

3 Psychopathology and AI Safety

Since its inception, AI has been closely connected to psychology and cognitive sciences [7]. This connection flows in both directions: AI researchers study biological cognition and behavior as inspiration for engineered intelligence, and cognitive scientists explore AI as a framework for synthesis and experimental analysis of theoretical ideas [5]. An instance of this interconnection is Reinforcement Learning (RL), where the computational algorithms of RL, such as Temporal Difference (TD) learning were originally inspired from the dopamine system in biological brains [14]. On the other hand, the work on TD learning has provided mathematical means of modeling the neuroscientific dynamics of dopamine cells in the brain, and has been employed to study disorders such as schizophrenia and the consequences of pharmacological manipulations of dopamine on learning [12].

With regards to the relationship between psychological disorders and AI safety, there are scarce and sparse resources available in the literature. Recent papers by Ashrafian [1] and Yampolskiy [15] [17] present high-level arguments for the existence and emergence of mental disorders in AI. One such argument presented in [1] is based on the analogy of David Chalmers' philosophical zombie (p-zombie). In this analogy, the p-zombie is considered to be a fully functioning robot that acts exactly like a human-being, which is not necessarily equipped with vague notions of consciousness [17]. The fact that this robot is capable of acting indistinguishably from humans is then used to justify that it is also prone to developmental and cognitive abnormalities that lead to misbehavior and anomalous cognition.

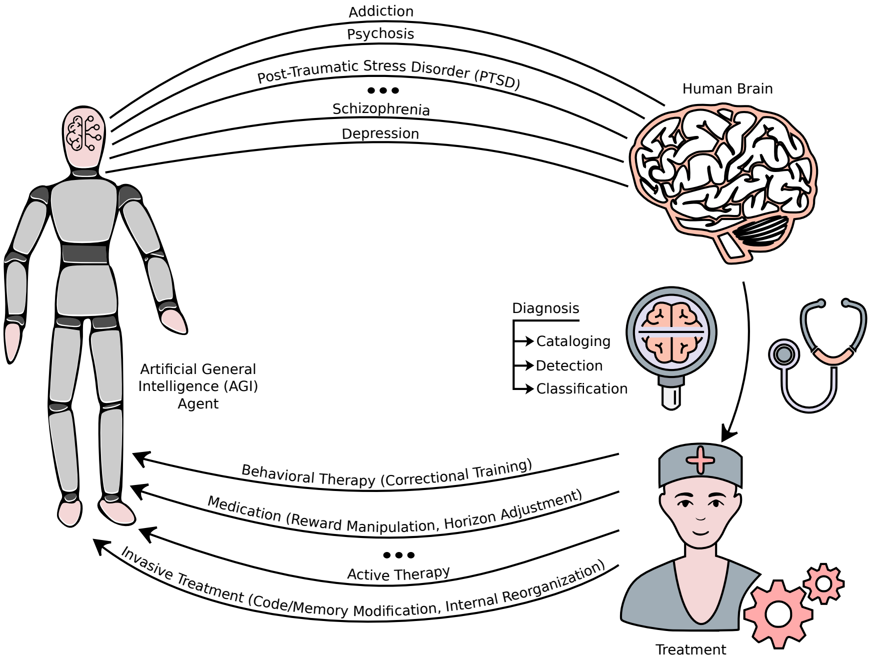
Addiction
Psychosis
Post-Traumatic Stress Disorder (PTSD)
•••
Schizophrenia
Depression
Human Brain
Artificial General Intelligence (AGI) Agent
Diagnosis
Cataloging
Detection
Classification
Behavioral Therapy (Correctional Training)
Medication (Reward Manipulation, Horizon Adjustment)
•••
Active Therapy
Invasive Treatment (Code/Memory Modification, Internal Reorganization)
Treatment

Fig. 1. A psychopathological approach to safety engineering in AI and AGI.

Furthermore, many aspects of failures in AI safety can be viewed as psychological disorders. For instance, wireheading in AI can manifest as delusional and addictive behavior. [15]. Similarly, sequences of interactions with extremely negative rewards and stresses within the exploration/exploitation trajectories of RL-based AI can potentially give rise to behavioral disorders such as depression and Post-Traumatic Stress Disorder (PTSD) [1]. Furthermore, the generic manifestation of the value alignment problem [11] in AI is in the form of behavioral characteristics that are harmful to either the agent or the environment and society, which falls well within the definition of psychological disorders.

While [1] and a few other papers (e.g., [3][11]) present high-level arguments on the advantages of investigating the psychopathology of AI, there remains a wide gap in satisfying the need for technical studies and practices. This paper presents a research agenda that will fill this gap via the following proposals, also illustrated in Figure 1.

4 Directions of Research

Developing solid grounds for research on the Psychopathology of AI requires investigations in three main areas: Modeling and Verification, Diagnosis, and Treatment. In this section, we define and discuss the scope of each area.

4.1 Modeling and Verification Tools

While the descriptive similarities of human psychopathology and AI failures provide some insights into adopting such abstractions for AI safety, taking an engineering approach requires formal and mathematical modeling of the aspects and dimensions of these similarities. Such formalisms may benefit from those that already exist in the realm of cognitive and medical sciences, such as cognitive architectures [10] and RL-based models of the dynamics in mental disorders (e.g., [12]). Also, the quantitative analysis of such disorders necessitate the exploration and development of new models of AI and AGI based on such paradigms as neuroeconomics, complex adaptive systems, control theory, and dynamic data-driven application systems.

Furthermore, verification and validation of such models and the ensuing theories requires the development of experimental frameworks and simulation platforms. Such platforms must provide the means for wide ranges of experiments on emergence and dynamics of behavioral and cognitive disorders in arbitrary and context-dependent scenarios, and shall be compatible for various agent and environmental models.

4.2 Diagnosis and Classification of Disorders

This venue is on investigating and development of techniques for diagnosis of disorders in AI. Within the context of AI safety engineering, diagnosis refers to two inter-related tasks: first is to detect anomalous behaviors, and the second is to classify the type of anomalous behavior as a first step towards treatment. Detection of undesired behavior is an active topic of research in AI safety, with initial solutions such as tripwires and honeypots [11] already proposed and investigated. We propose to extend current state of the art in detection through adoption and automation of parallel techniques in psychopathology. Similar to diagnostics criteria in human psychology [2], a promising approach is to identify statistical deviations in behavior, as well as general indicators of misbehavior. To this end, development of machine learning approaches similar to those applied in cybersecurity for threat and intrusion detection can be a promising direction. Furthermore, generic indicators of misbehavior can be learned from models trained on simulated and annotated scenarios of disorders.

Once a misbehavior is detected, the next step is to characterize and classify the disorder that has led to such behavior. A prerequisite to this process is having a catalog of different disorders and the corresponding criteria for diagnosing such disorders. Therefore, a necessary step is the compilation of representative and experimentally verified disorders, such as addiction and anxiety in RL agents, along with manually and automatically generated criteria and characteristics of each disorder based on behavioral observations. This task shall aim to produce human- and machine-readable catalogs as AI analogues of APAs DSM 5 [2].

Besides general behavioral characteristics, there are other sources of data that can be of diagnostic value. Instances include indicators of disorders that are obtained through direct and targeted interactions with AI (similar to psychiatric evaluation of human patients), non-invasive analysis of internal states and parameters (similar to F-MRI and EEG tests of human patients), and induction or invocation of internal debug modes (similar to states of hypnosis). Exploring such ideas and approaches may greatly enhance the accuracy of diagnosis, and lead to novel techniques for psychoanalysis and diagnostics of AI and AGI.

4.3 Treatment

When a disorder is diagnosed in an AI agent, it is not always feasible to simply decommission or reset the agent. In such cases, it is often preferable to pursue treatment via minimally destructive techniques that correct the misbehaviors of agent, while preserving the useful traits learned by that agent. Such treatments need to satisfy a number of challenging requirements. Advanced AI are complex adaptive systems, and therefore minor perturbations of one component may lead to unintended consequences on local and global scales. For instance, correcting a developmental disorder by removing a series of harmful experiences from the memory of an AI may lead to behavioral changes that are even more undesirable than the original misbehavior. Therefore, effective treatments must either be minimally invasive or non-invasive at all.

Inspired by psychopathological parallels, we propose two general approaches to treatment of pathologies in AI. One is correctional training, which adopts the approach of behavioral therapy. This approach is to retrain an agent in controlled environments and scenarios, such that harmful experiences can be remedied or alleviated through new experiences. The second approach parallels that of medication therapy, in which the reward signals of AI agents are artificially manipulated via external means to adjust their behavioral policies. This is similar to the use of anti-depressants and anti-psychotics in treating disorders related to production and inhibition of dopamine and serotonin in human brains.

5 Conclusion

This paper presents the argument that while current research in AI safety is generally focused on design and mitigation problems, the complexity of AGI will render such analysis as difficult as those that capture biological intelligence and disorders. Hence, studying the complex dynamics and mechanisms of emergent failures in AI and AGI can greatly benefit from abstractions that parallel those of biological intelligence. Considering the practical objectives of diagnosing and treating misbehaviors in AGI, we propose that psychopathological approaches provide tractable settings while benefiting from various bodies of knowledge. Accordingly, we present a high-level research agenda that includes explorations of parallels between human and AI psychopathology, development of methodologies for diagnosis of behavioral pathologies in AI, and propose techniques for treatment of such disorders.

As the paper detailes, psychology and AI enjoy a bi-directional flow of inspirations. A major impact of the proposed research is the production of outcomes that can be of use and inspiration to current research in psychopathology and cognitive sciences. Furthermore, the results of this work may provide a deeper understanding of the safety requirements and guidelines for designing advanced AI and AGI, while guiding policy makers on the risks and potential solutions involved in the integration of AGI into societies. We hope that this paper motivates initial efforts in laying solid foundations for future research and developments in this scarcely explored but promising venue.

References

1. Ashrafian, H.: Can artificial intelligences suffer from mental illness? a philosophical matter to consider. Science and engineering ethics **23**(2), 403–412 (2017)
2. Association, A.P., et al.: Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub (2013)
3. Atkinson, D.J.: Emerging cyber-security issues of autonomy and the psychopathology of intelligent machines. In: Foundations of Autonomy and Its (Cyber) Threats: From Individuals to Interdependence: Papers from the 2015 AAAI Spring Symposium, Palo Alto, CA. http://www. aaai. org/ocs/index. php/SSS/SSS15/paper/viewFile/10219/10049 (2015)
4. Butcher, J.N., Hooley, J.M.: Apa handbook of psychopathology: Psychopathology: Understanding, assessing, and treating adult mental disorders, vol. 1 (2018)
5. Collins, A., Smith, E.E.: Readings in cognitive science: A perspective from psychology and artificial intelligence. Elsevier (2013)
6. Davis, T.: Conceptualizing psychiatric disorders using four ds of diagnoses (2018)
7. Dennett, D.C.: Artificial intelligence as philosophy and as psychology. Brainstorms: Philosophical essays on mind and psychology pp. 109–26 (1978)
8. Kelly, J., Gooding, P., Pratt, D., Ainsworth, J., Welford, M., Tarrier, N.: Intelligent real-time therapy: Harnessing the power of machine learning to optimise the delivery of momentary cognitive–behavioural interventions. Journal of Mental Health **21**(4), 404–414 (2012)
9. Kendler, K.S.: The dappled nature of causes of psychiatric illness: Replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. Molecular psychiatry **17**(4), 377 (2012)
10. Kotseruba, I., Tsotsos, J.K.: A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. arXiv preprint arXiv:1610.08602 (2016)
11. of Life Institute, F.: The Landscape of AI Safety and Beneficence Research: Input for Brainstorming at Beneficial AI 2017. Future of Life Institute (2017)
12. Montague, P.R., Hyman, S.E., Cohen, J.D.: Computational roles for dopamine in behavioural control. Nature **431**(7010), 760 (2004)
13. Nordström, A.L., Farde, L., Wiesel, F.A., Forslund, K., Pauli, S., Halldin, C., Uppfeldt, G.: Central d2-dopamine receptor occupancy in relation to antipsychotic drug effects: a double-blind pet study of schizophrenic patients. Biological psychiatry **33**(4), 227–235 (1993)
14. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
15. Yampolskiy, R.V.: Utility function security in artificially intelligent agents. Journal of Experimental & Theoretical Artificial Intelligence **26**(3), 373–389 (2014)
16. Yampolskiy, R.V.: Taxonomy of pathways to dangerous artificial intelligence. In: AAAI Workshop: AI, Ethics, and Society (2016)
17. Yampolskiy, R.V.: Detecting qualia in natural and artificial agents. arXiv preprint arXiv:1712.04020 (2017)