Chapter 2
Predictability of AI Decisions

Grzegorz Musiolik
https://orcid.org/0000-0001-7282-4760
German Aerospace Center, Germany

ABSTRACT

Artificial intelligence evolves rapidly and will have a great impact on the society in the future. One important question which still cannot be addressed with satisfaction is whether the decision of an intelligent agent can be predicted. As a consequence of this, the general question arises if such agents can be controllable and future robotic applications can be safe. This chapter shows that unpredictable systems are very common in mathematics and physics although the underlying mathematical structure can be very simple. It also shows that such unpredictability can also emerge for intelligent agents in reinforcement learning, especially for complex tasks with various input parameters. An observer would not be capable to distinguish this unpredictability from a free will of the agent. This raises ethical questions and safety issues which are briefly presented.

INTRODUCTION

Artificial Intelligence (AI) is the science of automating intelligence. This includes many factors like reasoning, memorizing or speaking and calculating. If intelligent agents got close to human cognitive abilities one day they would make decisions on their own. This chapter addresses the question of how predictable these decisions can be. AI is often thought to be a recently established field of research but the concept dates back to the 1950s. The chapter shows the historical evolution and gives examples of the current state of research on AI reaching from medicine to spaceflight. In a second step, chaotic systems in mathematics and physics are introduced. The main property of such systems is the unpredictable change of their evolution when varying the initial conditions only slightly. In popular science, this effect is often referred to as the butterfly effect. The chapter gives an overview over the historical evolution of chaotic systems and shows some particular examples. The chaotic behavior of such systems is then transferred to the decision making process of AI. The chapter shows that in some cases the AI behavior cannot be predicted and statistical methods should be used instead. The main objectives of the chapter are:

DOI: 10.4018/978-1-7998-3499-1.ch002

1. Provide an introduction to the historical and recent developments in AI-research based on various examples from physics and life sciences.
2. Introduce chaotic behavior in mathematical and physical systems.
3. Illustrate analogies between chaos in mathematics and physics and AI and show that AI-decisions cannot be predictable in some cases.
4. Propose a solution to the AI unpredictability by creating a statistical model addressing the risks of the outcomes following a decision.

BACKGROUND

The beginning of the information age is the foundation of AI. Ever since the introduction of the information theory by Shannon (1948) who describes the informational entropy and the computational theory by Turing (1936) the possibility of an artificial brain seems natural. McCulloch and Pitts (1943) point out that the neuronal activity can possibly be described by computations. The term "artificial intelligence" is specified only a few years later at the Dartmouth conference in 1956 followed by a vast development in the field of AI. One of the most important milestones at this time dates back to the concept of the Perceptron by Rosenblatt (1957). This concept is the origin of the connectionism which tries to explain cognitive phenomena by an artificial neural network. Rosenblatt believes the Perceptron of being able to learn and make own decisions. However, Minsky and Papert (1969; 2017) describe fundamental concerns related to the Perceptron such as the difficulty of handling a XOR-function which finally leads to the first AI-winter. The low computational power is another significant practical limitation for intelligent algorithms at this time. Further major developments in AI research do not take place until the early 1980s. One of the following breakthroughs is the commercialization of "expert systems" which are able to help an operator solving complex problems based on if-then relationships. An expert system can e.g. be used for medical diagnostics as described by Weiss et al. (1978). Around the same time, Hopfield (1982) describes the Hopfield network which is a recurrent artificial neuronal network with binary nodes. This network is able to learn and can be used to model the human associative memory. Another important work on the backpropagation for artificial neuronal networks is done by Rumelhart et al. (1986). These developments finally set the course for machine learning as we know it today.

AI evolves rapidly and will likely change many aspects of our society in the near future. There are countless applications which comprise big data analysis based on machine learning algorithms. For example, such algorithms are helpful to find the shortest navigational path between two points in space (Ikeda et al., 1994). With an increasing traffic density machine learning algorithms could help to navigate more efficiently. Furthermore, learning algorithms for self-driving cars could be able to take over the human control over the vehicle completely and bring advantages for the infrastructure (Bojarski et al., 2016; Daily et al., 2017). A complex application for artificial intelligence is the health care (Ramesh et al., 2004; Ellahham et al., 2019). Intelligent agents can be as good in analyzing medical image data as physicians (Beam and Kohane, 2018). A particular example is the enhanced diagnosis in cardiac imaging (Dilsizian and Siegel, 2004). In future, AI-algorithms may control surgery robots and enhance medical standards in this field (O'Sullivan et al., 2019). These medical developments may save more patients and take the heat off medical personal in future.

Another very popular application of AI is the self-learning of games like Chess, Shogi or Go (Silver et al., 2018). AI-algorithms are implemented to analyze a large amount of games and learn strategies