Chapter XLI
From Coder to Creator:
Responsibility Issues in Intelligent Artifact Design

Andreas Matthias
Lingnan University, Hong Kong

ABSTRACT

Creation of autonomously acting, learning artifacts has reached a point where humans cannot any more be justly held responsible for the actions of certain types of machines. Such machines learn during operation, thus continuously changing their original behaviour in uncontrollable (by the initial manufacturer) ways. They act without effective supervision and have an epistemic advantage over humans, in that their extended sensory apparatus, their superior processing speed and perfect memory render it impossible for humans to supervise the machine's decisions in real-time. We survey the techniques of artificial intelligence engineering, showing that there has been a shift in the role of the programmer of such machines from a coder (who has complete control over the program in the machine) to a mere creator of software organisms which evolve and develop by themselves. We then discuss the problem of responsibility ascription to such machines, trying to avoid the metaphysical pitfalls of the mind-body problem. We propose five criteria for purely legal responsibility, which are in accordance both with the findings of contemporary analytic philosophy and with legal practise. We suggest that Stahl's (2006) concept of "quasi-responsibility" might also be a way to handle the responsibility gap.

INTRODUCTION

Since the dawn of civilization, man has lived together with artifacts: tools and machines he himself has called into existence. These artifacts he has used to extend the range and the quality of his senses, to increase or replace the power of his muscles, to store and transmit information to others, his contemporaries or those yet to be born. In all these cases, he himself had been the controlling force behind the artifacts' actions. He had been the one to wield the hammer, to handle the knife, to look through the microscope, to drive a car, to flip a switch to turn the radio on

or off. Responsibility ascription for whatever the machines "did" was straightforward, because the machines could not act by themselves. It was not the machine which acted, it was the controlling human. This not only applied to the simple tools, like hammers and knives, but also to cars and airplanes, remotely controlled planetary exploration vehicles and, until recently, computers.

Any useful, traditional artifact can be seen as a finite state machine: its manufacturer can describe its range of expected actions as a set of transformations that occur as a reaction of the artifact to changes in its environment ("inputs"). The complete set of expected transformations is what comprises the *operating manual* of the machine. By documenting the reactions of the machine to various valid input patterns, the manufacturer renders the reader of the operating manual capable of effectively *controlling* the device. This transfer of control is usually seen as the legal and moral basis of the transfer of *responsibility* for the results of the machine's operation from the manufacturer to the operator (Fischer & Ravizza, 1998). If the operation of a machine causes damage, we will ascribe the responsibility for it according to who was in control of the machine at that point. If the machine operated correctly and predictably (that is, as documented in the operating manual), then we will deem its operator responsible. But if the operator can show that the machine underwent a significant transformation in its state which was not documented in the operating manual (e.g. by exploding, or failing to stop when brakes were applied) then we would not hold the operator responsible any longer, and precisely for the reason that he did not have sufficient *control* over the device's behaviour to be able to assume full responsibility for the consequences of its operation.

With the advent of *learning, autonomously acting* machines, all this has changed more radically than it appears at first sight. Learning automata, as we will see, are not just another kind of machine, just another step in the evolution of artifacts from the spear to the automobile. Insofar as responsibility ascription is concerned, learning automata can be shown to be machines *sui generis,* in that the set of expected transformations they may undergo during operation cannot be determined in advance, which translates to the statement that the human operator cannot *in principle* have sufficient control over the machine to be rightly held responsible for the consequences of its operation.

Learning automata cause a *paradigm shift* in the creation, operation and evaluation of artifacts. In the progress of programming techniques from classic, imperative programming, to declarative languages, artificial neural networks, genetic algorithms and autonomous agent architectures, the manufacturer/programmer step by step gives up control over the machine's future behaviour, until he finds her role reduced to that of a *creator* of an autonomous organism rather than the powerful, controlling *coder* that she still is in popular imagination and (all too often) in unqualified moral debate.

In the course of this chapter, we will retrace the crucial points of this technological development. We will see how exactly the shift from coder to creator takes place and what this means for the problem of responsibility ascription for the actions of learning automata. It can be shown that the loss of control over the operation of such machines creates a "responsibility gap" which must somehow be bridged. Since humans cannot have enough control over the machine's behaviour to rightly assume responsibility for it, we will examine the question whether learning, autonomous machines could possibly be ascribed themselves responsibility for their own actions. We will discuss the prerequisites to machine responsibility and see that it does not necessarily mean that we will need to consider machines to be moral agents or even quasi-personal entities. Instead, responsibility ascription to a machine can be done without a shift in the metaphysical status of the machine using a "functional" approach to responsibility ("quasi-responsibility,"