International Journal of Technoethics
Volume 12 • Issue 1 • January-June 2021

Confounding Complexity of Machine Action:
A Hobbesian Account of Machine Responsibility

Henrik Skaug Sætra, Østfold University College, Halden, Norway

ABSTRACT

In this article, the core concepts in Thomas Hobbes's framework of representation and responsibility are applied to the question of machine responsibility and the responsibility gap and the retribution gap. The method is philosophical analysis and involves the application of theories from political theory to the ethics of technology. A veil of complexity creates the illusion that machine actions belong to a mysterious and unpredictable domain, and some argue that this unpredictability absolves designers of responsibility. Such a move would create a moral hazard related to both (a) strategically increasing unpredictability and (b) taking more risk if responsible humans do not have to bear the costs of the risks they create. Hobbes's theory allows for the clear and arguably fair attribution of action while allowing for necessary development and innovation. Innovation will be allowed as long as it is compatible with social order and provided the beneficial effects outweigh concerns about increased risk. Questions of responsibility are here considered to be political questions.

KEYWORDS

Attribution, Complexity, Hobbes, Instrumentalism, Responsibility

INTRODUCTION

How can we attribute praise, blame, and responsibility when machines perform actions? The question of machine responsibility and agency is an old one, but we are still seemingly confounded by the complexity of new technologies. When complicated machines act, so to speak, on their own, without their designers being able to control or predict and fully understand their actions, can they still be held responsible?

In this article the core concepts in Thomas Hobbes's framework of representation and responsibility are applied to the question of machine responsibility. This provides a simple and straightforward way of understanding the attribution of machine actions, and simultaneously narrows or eliminates the responsibility gap and retribution gap discussed in the literature on machine agency and responsibility (Danaher, 2016; de Jong, 2019; Gunkel, 2017; Köhler, Roughley, & Sauer, 2018; Nyholm, 2018; Tigard, 2020).

This account constitutes a challenge to modern approaches to machine responsibility, and in particular the view that modern machine complexity transcends traditional accounts of responsibility (Matthias, 2004). The challenge consists in taking us back to the basics to show that the basics are

DOI: 10.4018/IJT.20210101.oa1

not necessarily incapable of dealing with the actions of complex machines. Along the lines drawn by Köhler et al. (2018), Robillard (2018) and Tigard (2020), this article concludes that the gaps that causes concern might, in fact, be illusory, or simply the product of applying inappropriate frameworks for understanding the attribution of actions. In addition to this, the Hobbesian framework contains a distinction between natural and artificial agents, representation and responsibility of actions, and a general political framework that allows for the attribution of actions to machines for pragmatic reasons.

To determine whether artificial intelligence (AI) can bear responsibility, we must first understand what constitutes a *person*, *author* and an *actor*. The Hobbesian approach provides a way of avoiding much current confusion and controversy by relying on an instrumental theory of responsibility and accountability that does not fall foul of common objections to such an approach, such as stifling innovation (Gunkel, 2017). It is argued that stifling innovation is at times both necessary and legitimate, and that the question of what risks to accept in order to achieve innovation and economic growth, for example, is subject to political deliberation, as innovation and growth are only two amongst many goals of society. At the same time, the framework allows for the consideration of non-humans as artificial persons, if such an approach is deemed beneficial. Machines, then, could be assigned a form of personhood along the lines of limited liability corporations. The Hobbesian framework also shows how AI can be considered *artificial persons*, and if such a move creates gaps, these are ancient gaps.

First, the question of attribution of machine actions is examined, along with the Hobbesian framework of persons and representation. Secondly, the nature of modern machines is considered, as their complexity is claimed to constitute a fundamental challenge to traditional approaches to attribution of responsibility. Thirdly, the responsibility and retribution gaps are considered in light of the Hobbesian framework.

Hobbes's theory allows for a clear and arguably fair way of attributing machine action, while also allowing for necessary development and innovation. Responsible and beneficial innovation will be allowed as long as it is compatible with social order, and if the beneficial effects outweigh concerns about increased risk and moral hazard.

ATTRIBUTION OF RESPONSIBILITY

Modern machines are complex. They are so complex, in fact, that makers and operators of machines no longer understand them. Advanced machine learning and genetic algorithms are two examples of the techniques that are said to cause this (Matthias, 2004). This factor, some say, makes it unfair, unintuitive, or simply not right, to attribute responsibility for machine actions to machine makers or operators (Matthias, 2004). As emphasised by Tigard (2020), responsibility can entail attributability, accountability, or answerability. He employs a pluralistic account of *moral* responsibility and thus extends the analysis of the gaps beyond both law and questions of accountability. Accountability is the main focus of this article, as will become clear when the Hobbesian framework of representation is presented.

Attributing the actions of machines to humans is also associated with negative consequences, as it could stifle innovation and prevent beneficial use of new technologies (Gunkel, 2017; Matthias, 2004). It could even deprive people of their perceived need for retribution (Danaher, 2016). In discussing the *gaps* thus created between who *has* responsibility and blame and whom we attribute it to, de Jong (2019) argues that the complexity of the *production* of modern technology is yet another nail in the coffin for what she labels "traditional approaches" to attributing responsibility.

In this article it is argued that the traditional approach *is* still viable, and that objections to its use derive mainly from the confounding complexity of new technologies. The account here presented is an *instrumentalist* account based on the view that, when it comes to responsibility, machines are tools under human responsibility. The machines essentially "assist the animate being" in the realisation of the goals and pursuits of others (usually human beings) (Sacksteder, 1984).