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).

This view is similar to the views of Koops, Hildebrandt, and Jaquet-Chiffelle (2010), Bryson (2010), Calo (2015), Nyholm (2018) and (Tigard, 2020). They all argue in favour of traditional approaches involving the attribution of responsibility to human beings, and it will here be shown not only that these approaches can handle complexity, but also that the complexity itself is somewhat illusory. The goal of the article is to highlight the nature of the gaps and to show that old theories are fully capable of dealing with the attribution of machine actions. Furthermore, while some of the instrumentalist accounts reject the possibility of limited machine agency, the account here developed shows how robots *can* be considered limited agents, without necessarily being able to own their own actions. Also, rather than emphasising *agency* and the moral aspects of responsibility, it enables us to treat these questions as political questions informed by, but not determined by, philosophical considerations of *moral* agency.

Thomas Hobbes (1588–1679) was a political theorist much concerned with social order and political legitimacy. Hobbes lived in a time of great conflict, and today he is most famous for his systematic approach to political philosophy, which led him to prefer the rule of few and to emphasize that the sovereign must have absolute authority in order to be able to provide security and order (Hobbes, 1946). Using the metaphor of a *social contract,* Hobbes argued that a government with absolute power could be legitimate. Our worst fear, he argues, is to live in fear of violent death, and civil war is the condition most conducive to produce such fear. To avoid this condition, Hobbes imagines that we would be willing to transfer our natural right of liberty to a sovereign, who would in turn allow us all the liberty that is compatible with our own protection. While he acknowledges that a sovereign with absolute power is not ideal, he argues that the downsides are less weighty than the downsides of limited power prone to produce instability and conflict (Hobbes, 1946).

Hobbes' use of a contract as his mechanism for legitimate transfer of rights and power underscores that the ideas of representation and responsibility are important parts of his philosophy. For example, he discusses in detail how the social contract allows the sovereign to act as a representative of the individual citizens, with the possibility that a citizen is acted upon with his *own* authority (Hobbes, 1946). While the instrumental approach mainly involves seeing the human being as the actor, this Hobbesian notion of authorising representatives to act *may* also be applicable to machines. This would open up for a more nuanced approach to machine action than that of, for example, Robillard (2018), who rejects *any* form of machine agency (unless they have become proper moral agents).

While Hobbes's focus on representation and responsibility in itself makes him an interesting theorist for the question at hand, the fact that he examines these issues in the case of an *artificial being* makes it even more relevant. Haugeland (1989) refers to Hobbes as the "grandfather of AI" and the computational mind, but it is not Hobbes' mechanism in itself that is of most interest when discussing these issues. Instead, it is the fact that he considers the commonwealth an "artificial man*"*. When human beings contract and create new constellations of rights and responsibilities, this, Hobbes argues, "resemble[s] that fiat, or the Let us make man, pronounced by God in the Creation" (Hobbes, 1946). Human beings can create a commonwealth, and they can also turn their own creations – machines of various kinds – into artificial persons, when this is deemed beneficial or necessary. This creative potential of human beings is emphasised by Sacksteder (1984), who writes of man "the artificer" – the "mechanic, who contrives machines".

Hobbes is here used because of his concept of authority and representation, his theory of punishment, and his idea of the social contract as the basis of government legitimacy. The result is a *Hobbesian* (inspired by Hobbes) account of the attribution of machine action. The account highlights the illusory nature of both the responsibility gap and the retribution gap, and it is an instrumental and pragmatic theory that does *not* suffer from some of the common objections to instrumentalism, such as the stifling of innovation and the creation of gaps.

Computer Programs and Artificial Agents

If we fast-forward more than 350 years from Hobbes's heyday to the present age, we find ourselves in an era during which an *artificial agent* tends to evoke the idea of a computer-based entity capable of performing certain actions. The purpose of this article is to examine the degree to which these modern artificial agents are amenable to Hobbes' older framework for regulating and understanding the acts of artificial persons. The analysis begins with the question: *Who defeated Lee Sedol in the five-game Go match in Seoul in March 2016?*[1] Go is an ancient board game, and Lee Sedol was one of the best at playing it. In 2016, however, he was defeated in a game involving a computer.

In order to start the analysis Sedol's *apparent* opponent must be identified: Google's AlphaGo (Google, 2020a). Who, or what is AlphaGo? The choice between these two words – who/what – immediately reveals that the conceptual waters we traverse are rough. *Who* refers to a living person like ourselves, whereas *what* refers to a thing (Gunkel, 2019).

On a basic level, AlphaGo is a *computer program* developed by DeepMind (Google, 2020b). But what *is* a computer program, and can it be ascribed agency and the ability to play and win games? A computer *program* is created by writing code in a programming language. The code is then run through a compiler, which creates executable files. These files are run by a compatible operating system, which in turn runs on a *computer.* The computer running the application is, in a physical sense, the *machine* that is involved in playing.

An autonomous vehicle involves some additional aspects as well. Code is produced, compiled, and run by a computer; this time it is located in a mobile shell. This mobile shell can even propel itself. It also *senses* its environment and attempts to react to the stimuli received in order, for example, to avoid running over pedestrians. The machine that is an autonomous vehicle is thus a *robot* according to common definitions of the term (Brooks, 1991).

Thus far, we have a computer, running an operating system, running compilers which can turn code into executable files, which in turn become a computer program like AlphaGo or the software used to control autonomous vehicles.

Another creature is also involved in the process, however. The code run through the compiler is created by *human beings*.[2] This could be the DeepMind team, or any other programmer. The computer, with its programming environment, compiler, and ability to run programs *enables* people to run a computer program. Alternatively, we could say that the humans *command* the computer to run AlphaGo.

This takes us to the central questions, where the Hobbesian notions of responsibility and representation are used to answer questions such as, are humans the authors of programs, and thus also of the actions that follow? Is the compiler a co-author, and are the operating system and the computer enablers? Nyholm (2018) introduces the idea of *collaborative* agency, and thus considers such ideas seriously. However, such concepts may not be necessary once we remove the veil of complexity that obscures modern machines.

Hobbesian Persons and Owners of Actions

The veil shall be removed by way of applying a Hobbesian framework of representation. While Pitkin (1967) argues that Hobbes provided the first systematic English account of political representation, Skinner (2005) convincingly argues that this might be an exaggeration. Nevertheless, Hobbes provides an early influential account of representation, which is of relevance here because it entails "the act of speaking or acting in the name of someone else, and more specifically doing so with permission or authority" (Skinner, 2005). Such an account of representation allows for (a) straightforward attribution of action by examining the origin of an act, and (b) asking whether machines might in fact be *representatives* of human beings, instead of mere tools.

A person is defined as the origin of actions or words, either his own, or those of others (real or fictional). If we consider the person to be capable of owning his actions, or words, we call them a *natural person* – other persons are *feigned* or *artificial* (Hobbes, 1946). A natural person is an active

cause, Sacksteder (1984) argues, and can thus own actions and authorise acts by others. Koops et al. (2010) discuss similar concepts that have been used to examine the possibility of *legal* personhood for new entities. The debates are there traced back to the early 1990's. The Hobbesian framework here presented goes a lot further back, however, and discussions and about Hobbes's mechanism and automata are highly relevant for seemingly new phenomena such as robots and advanced AI (Sacksteder, 1984).

A person sounds quite human, so let us distance our actors somewhat from the human by returning to the roots of the word. While *persona* in Latin refers to the "disguise, or outward appearance of a man, counterfeited on the stage", in Greek we have *prosopon*, meaning face, for the same function (Hobbes, 1946). In this sense, non-humans might be personas, or faces. Gunkel (2018) expounds upon the philosophy of Levinas and the concept of face and others, as he argues that machines might also have such roles. If machines are natural persons, this could indeed mean that old theories were vulnerable to the gaps here discussed. Furthermore, if this is the case, the need to seriously consider the *rights* of machines would also emerge (Gunkel, 2018; Robillard, 2018). However, even if machines are considered to be Hobbesian artificial persons, this does not imply that a gap exists, as will be shown. Koops et al. (2010) further states that legal *personhood* is a mask that serves the purpose of separating legal entities and "physical persons or other entities". This highlights the *role* a machine might play by being considered an artificial person. Pettit (2001) also discusses the etymology of persons and the self, and refers to the Hobbesian notion that the concepts are restricted to those that can speak and think of themselves by "first-person indexicals". The importance of "being able to give expression" of beliefs and desires matters, he states. In the Hobbesian framework here developed it is not argued that machines have *own* states of belief or desire, but their ability to express such states in *others* are considered important with regard to the role they can play as artificial persons.

One potential way of viewing such entities is to liken them to limited liability corporations – "legal person distinct from owners and directors, as well as a separation of ownership and control" – the creation of which Hobbes might himself have been involved in (Jessen, 2012). The idea of limited liability corporations might lead to the consideration of limited liability *machines*, as will be discussed later.

Following these conceptualizations, we can say that to *personate* is to *represent* oneself or someone else. Those who act can thus do so as themselves, or as another, as a representative, or actor (Hobbes, 1946).

The question remains, then, whether machines can be *natural* persons. If so, this will enable them to own actions. If they cannot be *natural* persons, they might still in some sense be *artificial* ones, as they can both convey words and perform certain actions. In this sense, they are clearly distinct from the most basic objects and tools. An artificial person, however, is an *actor*, while there is some other *owner* of the actions. These owners are natural persons. An artificial actor that has been provided with the *right* to perform some action, has *authority* (Hobbes, 1946). This account is thus different from that Robillard (2018), who argues that a robot quite simply either *is* a moral agent or not – a question that is not pursued in depth here.

In a legal sense, however, machines need not be considered persons at all. Koops et al. (2010) provide an extensive account of accountability and machine actions, in which they also discuss such questions as giving machines personhood of various kinds. Their account thus extends the theoretical debates in this article into the legal domain – a domain Tigard (2020) opted not to discuss as he emphasised a pluralistic moral approach. The purpose of the current article, however, is to highlight the illusory nature of the consequences of machine complexity, while also providing a deeper theoretical framework for analysing questions such as retribution and the role of the *political* domain when it comes to trade-offs between, for example, innovation and safety.

This takes us to the application of the Hobbesian account of machine responsibility. For our example of the Go game, we need to determine whether or not AlphaGo is the *author* or the *representative* of an author of the games played. Or neither. If AlphaGo is an artificial person – a

representative of some person who is the author of the moves it plays – the actions of AlphaGo are the actions of this other person. The question, then, is whether a machine can own the actions it performs, and be a natural person, or at least a corporation-like limited liability person. Answering these questions involves coming to grips with the complexity of modern machines.

THE CONFOUNDING VEIL OF COMPLEXITY

The main argument of this article is that we are confounded by complexity. This refers to the idea that it somehow matters whether humans can foresee the actions of the machines they make, as "automated systems are making decisions that cannot be fully controlled or predicted" (de Jong, 2019). The machines are, Matthias (2004) states, unpredictable *in principle* for their creators. However, this (a) is not true, and (b) would not matter.

There are certain key technologies involved in the creation of AlphaGo, and *machine learning* and *genetic algorithms* seem to be particularly to blame for the confusion that has arisen (Matthias, 2004). AlphaGo is not taught to play the game of Go well by human experts of the game. Instead, it is programmed to *learn* how to play well by human experts on machine learning.

The people involved in the DeepMind team could not foresee the moves AlphaGo would play, as these were well beyond even the capacity of top human players' understanding of the game (Metz, 2016). This is arguably one cause of the responsibility gap some perceive. While the gap is often discussed in terms of culpability, here it involves the attribution of praise. If the mechanisms behind the success of DeepMind are not properly understood, it somehow seems wrong to state that someone who is not an expert in the game of Go should be given credit for beating one of its top champions by way of a machine.

However, it is wrong to say that the actions of AlphaGo are *in principle* unpredictable. They are *practically* unpredictable, because of limited human cognitive powers. But if we could imagine a human being with superior cognitive powers, such a being *could* understand the machine learning processes and predict the outcomes, *in principle.* Robot actions are "fully constitutive of the programmer's decisions and intentions" (Robillard, 2018). This is similar to the view of automata as the manifestation of its makers purposes and goals. They have no intent, no will, and their "movements are not [their] own" and these movements are not mystical or disconnected from human responsibility (Sacksteder, 1984). Saying otherwise turns AI into something *mysterious.* But "AI is not magic" – it is the application of known principles of mathematics, statistics and engineering (Marcus & Davis, 2019). However, even if a machine does not *understand* anything, and is neither mysterious nor magical, it *can* to a certain degree act, and there is nothing, in theory, preventing us from providing it with limited legal personhood in accordance with such abilities, according to Solum (1991).

The remaining question, then, is: *does it matter that a machine is unpredictable?* An example can help in answering this. Say person *X* writes a computer program that each minute randomly chooses between the number 0 or 1. *X* then attaches the machine to a real gun and aims it towards the street outside their house. Whenever the machine chooses 1, the trigger is pulled, and a bullet is fired. There are usually few people walking by *X*'s house, but one day a police officer shows up at their door, wondering what happened to cause the death of the person on the pavement outside their house.

"Oh, that must have been my machine!", *X* would say.

"I see," the police officer states in surprise, "You should come with me, then."

"But it was not *me", X* replies, "I had no idea when it would shoot."

The intent of the example should be clear by now. *X* has made a machine that could *not* be fully predicted or controlled once they had decided to turn it on and deploy it. But does this, in any meaningful way, absolve *X* of responsibility, or blame? It does not, and the example illustrates the confounding complexity of AI and machine learning. Although these technologies are more complex than that of a random killing machine, the principles involved are exactly the same. We can use some set procedure for identifying the human being deserving responsibility, such as the one proposed

by Nyholm (2018). This is similar to the arguments of Köhler et al. (2018), who argue for human responsibility in such cases. An example involves the application of autonomous weapons, where some human is considered to be the *commander* responsible – the one in control of the *application* of the weapons, but not in practical control of every action of the sophisticated weapons.

One important aspect to consider is the incentives created by our rules and regulations. If we state that a machine (or any other form of) unpredictability absolves an actor of responsibility, we may be creating great incentives for *not* creating transparent and explainable AI. In the *Ethics Guidelines for Trustworthy AI*, *transparency* is considered one of the core requirements for achieving trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019). One aspect of transparency is *explainability,* which entails a requirement that "the decisions made by an AI system can be understood and traced" by humans (High-Level Expert Group on Artificial Intelligence, 2019). Unpredictability would not absolve a creator of AI of responsibility if the criterion of explainability were made a legal requirement. However, such a move would most likely involve a trade-off that led to the creation of *less effective* and *more predictable* machines (High-Level Expert Group on Artificial Intelligence, 2019). A full discussion of such implications is beyond the scope of this article, but it suffices to note that if explainability is not formally required, there are clear incentives not to prioritize the development of explainable AI, and these incentives are further intensified by the claim that unpredictability absolves a creator of responsibility. Furthermore, it might also create the incentive to *state* that a machine is unpredictable and unexplainable, even if it may not be. This is a classic example of a situation in which absolving humans of responsibility for machine actions constitutes a moral hazard. The natural persons involved would then not bear the full cost of the actions of the machines, and they would thus be incentivised to *increase* the risk taken.

The Confounding Questions

The questions discussed here involve many different terms, most of which cannot be fully developed in this setting. Hobbes was very concerned about the confusion caused by such issues, and a reader of Hobbes is struck by his constant definitions of new terms, as he systematically builds his political philosophy (Hobbes, 1946). This is because language is a tricky beast, and words often lead us astray. Hobbes says words can be the "money of fools," who value them by the authority of others (Hobbes, 1946). He even goes so far as to list *speech* and *language* as fundamental causes of human conflict, because they allow us to represent "that which is good in the likeness of evil; and evil, in the likeness of good; and augment or diminish the apparent greatness of good and evil, discontenting men and troubling their peace at their pleasure" (Hobbes, 1946).

The *confounding complexity* of machine action is partly a result of the lack of proper definitions and uniform use of language. This is why some perceive gaps, for example, in the attribution of praise and blame (moral attribution) when machines perform actions. Many of the fundamental questions related to machine action sound simple, but they are in fact quite complex. To exemplify, if I run my robot mower, have I mowed the lawn? If I buy and start my Roomba, have I vacuumed?

Most would say that *X* pounded the nail into the board if *X* used a hammer. But does this change if *X* connects the hammer to the random triggering mechanism and then stepped away and watched the hammer shoot nails approximately every other minute? The chains of causality are more complex, and the distance from the action is increased, but nothing has *really* changed in terms of responsibility. Matthias (2004) argues that the *spatial* link between the action and the one responsible is broken, and Koops et al. (2010) aims to answer if the law is equipped to deal with such an increased physical distance between machines and those that employ them. The example of the random triggering mechanism shows, however, that the spatial aspect is of no importance for the attribution of responsibility.

The question that started it all was *Who beat Lee Sedol?* If *X* programmed AlphaGo, did *X* beat Lee Sedol in Go? Or is somehow *X*'s Go playing program suddenly a natural or artificial person capable of owning its own actions? There is nothing to suggest that it is. The complexity of the

machines is confusing because it allows individuals to use one skill (engineering/programming) to excel at *another* skill (i.e. Go-playing).

Saying that a human programmer with mediocre Go skills beat Lee Sedol in Go *feels* wrong, because they're not very skilled at playing Go. But this feeling does not matter, and attributing praise to them is appropriate – not because of their Go-playing skills, but because of their programming skills. This skill has become a *universal* skill that enables those who possess it to excel in most areas of human action.

It is akin to a physically weak person building an exoskeleton with immense power, allowing them to beat the world's strongest man in a competition of strength. Have they really then beaten him? In principle, yes, without question. There is no collaborative *agency* involved – no one with whom to share the glory – even if the exoskeleton did much of the literal lifting (Nyholm, 2018). The same goes for AlphaGo, which lifts intellectually; it is all the same, in principle. These questions, and questions regarding who is responsible for deaths caused by remote-controlled, or even autonomous, drones are the same as the question of who is responsible for the removal of dirt by a (mostly) manually operated excavator.

However, if the *rules* of the competition of strength *prohibited* exoskeletons, they would have won by *cheating*. And this is where the *rules of the game* come into play – the guidelines and best practices, but first and foremost: politics and the law.

RESPONSIBILITY AND RETRIBUTION

The first question involved in determining machine responsibility is whether machine complexity *changes* things or whether the Hobbesian concept of responsibility, and traditional notions of law, are capable of dealing with machines such as autonomous vehicles and programs playing games (Koops et al., 2010).

First, the law must describe a required process of diligence on the part of anyone who decides to deploy some machine in settings where they may impact other beings or the environment. Secondly, the law cannot proactively and specifically approve or decide what machines are allowed, unless one wishes to drastically change the dynamics involved in innovation and development. Matthias (2004) is correct in arguing that innovation should not be excessively disrupted. We must rely on general, strict laws of liability, and developers and entrepreneurs must be allowed to deploy new technologies subject to these rules. However, they must also be expected to adhere to strict norms of responsibility in development and entrepreneurship. In this respect, I support Calo (2015) and Nyholm (2018) and refer to them for more detail.

Instrumental accounts of responsibility will necessarily make developers cautious, as it reduces or eliminates the moral hazard associated with absolving developers of responsibility for cost of the risk related to their machine's errors and unpredictability. The chilling effect on innovation is one of the concerns addressed by Matthias (2004) and Gunkel (2017). However, while strict liability norms will make developers cautious, this is a good thing. The argument for less strict laws of liability seems to be based on the potential societal and economic benefits of more rapid innovation. This argument, however, is deeply problematic and requires careful consideration of the ethical foundations adhered to.

Pure utilitarianism might surely allow for reckless innovation at breakneck speeds, but there are alternatives to such an ethic – a Hobbesian ethic, for example, in which legitimacy is based on the idea that people create the sovereign in order to protect their primary interest, which is their own safety and survival. Once created, the government has a duty to protect these rights for *all* individuals and cannot simply apply utilitarian principles to sacrifice the safety of some so that others can acquire more wealth, for example. This is where the liberalism of individual liberty and the liberalism of free markets, innovation and progress, exhibit their quite different natures.

However, a Hobbesian framework *will* accept innovation that is responsible and beneficial for social order and prosperity. That is why we need the traditional framework of representation and

responsibility, as it allows for transparent, controlled and responsible development and implementation of new technologies.

Responsibility Gap

The responsibility gap is said to arise from the rationale that creators and operators cannot fully understand complex machines and thus cannot be held responsible for the effects of such machines. The gap is also said to result from a physical disconnect between humans and machines with varying degrees of autonomy (Matthias, 2004). Tigard (2020) relates this gap to "absent, unknown, or nonexistent sources of harm". However, the gap is illusory and caused in part by the veil of complexity.

There is no law that specifies that people cannot build a random killing machine, but it would still obviously be a crime to deploy it as was done in the example above. Building it is no crime, but deploying it *would* be. This is important for the question of attribution of responsibility. Level of *control* might not be as important. Let us assume that a person, *Y,* goes bow shooting in the middle of a crowded square and happens to miss the target and kill a bystander. An opponent, *Z*, uses a sniper rifle and manages to hit their target. Who would argue that *Y* is *less* responsible for their actions than Z would have been had they missed and shot someone, just because *Y* chose to use a device with littler accuracy – over which they had little control? It was *Y*'s *choice* to use the device in this setting, and they bear the full responsibility, despite their lack of control.

A more practical example involves two different autonomous vehicles. One is highly advanced and fully autonomous, while the other is simply equipped with a basic set of assistive features and safety features. Assume that these cars exist in a situation in which autonomous vehicles are not prohibited or highly regulated, and that the responsibility is determined by traditional traffic law. Person *A* uses his low-tech vehicle and causes no accidents. Person *B*, however, enables all the automatic features in his car and leans back in his seat to get some reading done. Their vehicle runs into a novel and confusing situation, and it so happens that it runs over a group of pedestrians. When the police arrive, the driver denies all responsibility. The car, he states, is so advanced that he had no idea what it would do, and his intention was only to get to work. The car *is* highly advanced, and when set free in a complex environment, its actions *are* practically unpredictable. However, using such a car is the equivalent of shooting with a highly inaccurate weapon in a public square, and the unpredictability involved implies no absolved responsibility.

Matthias (2004) also suggests that since the procedure of machine learning involves using errors as a method, we get a responsibility gap. This implies that people cannot be responsible for the errors that they cannot foresee. However, there is nothing controversial about saying that a person who *chooses* to use error as a method is clearly responsible for the consequence of such errors. When a machine fails, falls, errs, or does something that the maker did not intend, it will be a faulty machine, and not the machine it was intended to be. But the reason for this is the failure of design, and not the result of some mystical form of internal motion or magical agency in the machine, as an "automaton is moved according to the design built into it by the artificer", even if these designs are based on randomness, error, or unpredictability (Sacksteder, 1984). In a recent study it is shown that experts in the field of AI attributes responsibility to machines, in part due to the unpredictability of modern AI (Orr & Davis, 2020). While interesting, this shows that there may exist a *perceived* responsibility gap, and that people will *perceive* machines as responsible. Such phenomena do not create a *real* responsibility gap, however, just as the fact that people will become angry at, and maybe even strike, a door that they bumped their toe into, does not mean that the door is to blame.

Imagine an autonomous vehicle set free on the streets to learn to drive by trial and error. It seems clear that anyone who decided to do this is clearly liable for the consequences caused by this infant vehicle. For these reasons, innovative law and regulation must be examined, in order to allow for responsible innovation and development of AI. In Norway, for example, autonomous vehicles can be tested when deployed in limited ways, through the application of "regulatory sandboxes". These are already in place for transportation and financial technology, and a regulatory sandbox for

developing responsible AI solutions that respects privacy regulations is being developed (Ministry of Local Government and Modernisation, 2020; The Norwegian Data Protection Authority, 2020). Such mechanisms allow for (a) innovation and development, (b) limitation of public risk, and (c) the application of standard notions of responsibility and attribution of actions.

The rules of the game determine the attribution of responsibility. If playing Go with any sort of assistance is allowed, a player can win with the assistance of AlphaGo. If someone launches an autonomous vehicle, responsibility for accidents becomes a matter of examining (a) due process in development and (b) decision of how, when and where to deploy it. The modification of machines must naturally also be regulated, and this is an area of great importance for further research. Developers cannot always be held responsible for third-party modification or user "hacks", and this becomes increasingly relevant as the interplay between software and hardware, and for example open source software, is combined in machines deployed in society.

The *author* of an action is considered responsible for consequences in this framework. However, the real world is more complicated than this, and when complex chains of persons and things are involved in causing an outcome, the question of responsibility becomes difficult to disentangle, as de Jong (2019) argues.

There are several ways of attributing responsibility, and Carter (1999) distinguishes between four approaches: *Causal* attribution, *intentionality, moral* attribution or moral *culpability* (Carter, 1999). Tigard (2020) provides an account of the gaps based on a pluralistic conception of moral responsibility and has developed the implications of the different accounts in more detail. The current undertaking does not necessitate a full account of all the concepts, and they will be discussed only briefly to highlight the necessity of being explicit about which conception one adheres to.

Danaher (2016) prefers the *causal* approach, and thus opens for attributing responsibility to machines. This approach, however, is deeply problematic. First of all, where do we stop our investigations into the prior causes of an event (Sætra, 2019)? Furthermore, if *intentions* do not matter, all accidents can perfectly well be attributed to some person, even though this is not desirable for moral *or* legal reasons. Intentions should somehow matter, some might argue, but intentions are neither epistemologically available to us nor sufficient grounds for moral attribution. Roads paved with the best of intentions are positive contributions, but carelessness and recklessness can warrant liability despite good intentions.

Moral attribution is the approach that most fully captures all the required elements involved in the attribution of machine responsibility. It involves attribution of responsibility based on what we expect of an actor in terms of their due consideration of their own actions and acting in accordance with a certain set of norms. This relates to the *norms of responsibility* (Nyholm, 2018). I do not consider the criterion of moral *culpability*, which involves a moral evaluation of whether the actions performed were *right* or *wrong* (Carter, 1999).

In sum, Matthias's (2004) *starting* point – the attribution of responsibility to the creators or operators, according to how some machine was deployed, maintained, and the level of quality control, is still a workable concept. Nyholm (2018) provides a useful set of *responsibility loci* based on this traditional approach and demonstrates the political applicability of this approach. Furthermore, much work in responsible AI and robotics involve attempts to increase accountability, transparency, and the traceability of error in AI and robotic systems (Raji et al., 2020; Winfield & Jirotka, 2017). While such work is important, the argument here made is that a lack of transparency does not absolve a designer of duty. Rather, in order to further the development of transparency and explainability, designers of opaque original systems should be held fully accountable for any actions they cannot demonstrate not to result from their designs.

Retribution Gap

The responsibility gap, however, is not the only gap. Danaher (2016) introduces the *retribution gap*, which is the "mismatch between the human desire for retribution and the absence of appropriate

subjects of retributive blame". This is also discussed by Tigard (2020) as part of the *responsibility* gap, and he discusses, for example, anger, blame, and desire for retribution when two fictional lorries – one with a human driver, and one automated – runs over a child.

While interesting, a Hobbesian approach addresses both the concerns raised in Danaher's quote. First, the "human desire for retribution", and *retribution* itself, is *not* regarded to be the basis of punishment. Punishment, for Hobbes, is never about *retribution*. It is about correcting behaviour and providing beneficial incentives (Hobbes, 1946). In fact, much political philosophy is based on the need to *escape* from the effects of human needs for retribution. This aspect of human nature is important, but not for the reasons Danaher (2016) seem to imply. The experienced need for retribution cannot be the guide of policy regarding punishment, but must instead be tempered and downplayed. Danaher (2016) states that the responsibility gap arises from respecting the human need for retribution. Since this is not an aspect of Hobbesian justice, this gap is of little relevance in this context, even if it may be *true* in a descriptive sense.

Secondly, by assigning responsibility in the manner just described, there *are* subjects of blame: human beings; not machines that are assumed to be responsible because they arbitrarily happened to be positioned wherever our quest for causal factors ended. The retribution gap is partly the result of attributing agency and responsibility to machines. However, with the Hobbesian framework for assigning responsibility, there is no need for robot moral agency; not even *joint* or *collaborative agency* (Nyholm, 2018). These questions may be both morally and philosophically interesting, but with regard to the current question of *responsibility,* we need not consider robot moral agency at all; particularly if we remain in the realm of reality and the current status of AI (Marcus & Davis, 2019).

If an autonomous vehicle were judged responsible for accidents, and *punished*, many people would certainly find this both wrong and absurd. However, this is somewhat similar to the attribution of blame to other artificial persons: corporations (Jessen, 2012; Wellman, 2012). Corporations are at times held responsible for the actions of the people they comprise, and in this respect we can see a similar mismatch between the law and people's intuitive feelings of blame and retribution. A company might go bankrupt, but the people that are *felt* to be responsible may walk away scot-free.

It thus appears that the retribution gap is nothing new and is thus not necessarily specifically related to complex machines. If the traditional model of attributing responsibility to artificial persons, such as limited liability corporations, is seen to be the most desirable model for society, such a model is still compatible with the Hobbesian framework here described. People's *feelings* of injustice would only matter if they become so strong as to destabilise the system. People's need for retribution would thus be respected in the sense that it is dangerous. If their need for retribution is thwarted, they might reject the system and circumvent it. This, however, is quite different from arguing that the human need for retribution must be respected as *valuable*, and that it should be promoted and acknowledged fully in, for example, law.

Such a system can also accommodate limited liability machines, just as we have limited liability companies, if that is deemed the best way to preserve people's fundamental interests and rights. In this context, the political realism of Hobbes (1946) allows us to justify such trade-offs in a political setting. However, such a move would involve the creation of clear moral hazards. The questions of contract, personhood, and the legal domain are discussed in depth by Koops et al. (2010), who argue, broadly along the lines of the arguments made in this article, that it may make sense to consider machines "restricted" persons, but that *natural* personhood for machines does not seem to be a relevant solution for the foreseeable future.

CONCLUSION

The complexity of new machines entails that humans cannot "anticipate, completely control, or answer for" their actions, states Gunkel (2017). The first seems to be practically true, but the second is wrong if we also consider control in terms of development and deployment. This, in turn, entails

that the third is simply not correct. The argument made in this article is that the veil of complexity obscures the responsibility of human beings. When this veil is removed, it becomes apparent that modern machines are, in principle, not different from more primitive machines, which most people seem to believe our moral and legal theories can easily accommodate.

When it is argued that AI and robots should be responsible, this does not relate to *machine* responsibility. Responsible AI means that AI should be attributed responsibility just as little as transparent AI means that machines should be see-through. But in order for us to correctly attribute responsibility for machine actions, explainable AI would be beneficial. Despite this – explainable or not – the developers and/or operators of machines are responsible for their actions. They run the risk if they employ machines that are not transparent and explainable. Any other solution involves a clear moral hazard. We can certainly hope for some degree of responsibility on behalf of the makers of AI. However, should they act irresponsibly, our laws are in place to intervene. These are based in large part on principles similar to the ones in the Hobbesian framework presented here, and these principles are fully capable of dealing with complex machinery.

This framework is also capable of accommodating limited liability machines. This does not entail a consideration of robot capabilities, however, and we need not get into the demands of consciousness, intentionality, feelings, etc. necessary for a machine to be attributed some form of personhood (Koops et al., 2010). This would also not involve the "moral danger" of having to consider machine rights, etc (Robillard, 2018). These would be nothing more than legal fictions created for the sake of specific social benefits, just as limited liability corporations. If it is done, this is because debates in the political domain have led to the acceptance of some moral hazard and risk in order to achieve other political benefits, such as, for example, innovation and welfare. The Hobbesian framework is a realist one, in that such trade-offs can perfectly well be made, but it also entails a full theory of legitimacy and individual rights, which are parts of Hobbes's philosophy not covered here (Hobbes, 1946).

Complexity confounds, but it does not really change anything. When we remove the veil covering our modern machinery, we can see that they are not new magical creations capable of disrupting traditional notions of responsibility. They are both very useful and potentially very dangerous, and this implies that the human responsibility for such machines must be emphasised, not eroded. Even if machines should be granted legal personhood in some form, this is done in order to achieve societal benefits, and human responsibility for the machines remains intact – society would merely have decided to share it.

REFERENCES

Brooks, R. A. (1991). *Intelligence without reason.* A.I. Memo No. 1293.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (pp. 63–74). John Benjamin. doi:10.1075/nlp.8.11bry

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, 513–563.

Carter, I. (1999). *A measure of freedom*. Oxford University Press. doi:10.1093/0198294530.001.0001

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. doi:10.1007/s10676-016-9403-3

de Jong, R. (2019). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to nyholm. *Science and Engineering Ethics*, 1–9. PMID:31267376

Google. (2020a). *AlphaGo*. Retrieved from https://deepmind.com/research/case-studies/alphago-the-story-so-far

Google. (2020b). *Solve intelligence. Use it to make the world a better place*. Retrieved from https://deepmind.com/about/

Gunkel, D. J. (2017). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, 1–14. doi:10.1007/s10676-017-9428-2

Gunkel, D. J. (2018). *Robot rights*. MIT Press. doi:10.7551/mitpress/11444.001.0001

Gunkel, D. J. (2019). *How to Survive a Robot Invasion: Rights, Responsibility, and AI*. Routledge. doi:10.4324/9780429427862

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press. doi:10.7551/mitpress/1170.001.0001

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hobbes, T. (1946). *Leviathan*. Basil Blackwell.

Jessen, M. H. (2012). The State of the Company: Corporations, Colonies and Companies in Leviathan. *Journal of Intellectual History and Political Thought*, *1*(1), 56–85.

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability? In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral Agency and the Politics of Responsibility*. Routledge.

Koops, B.-J., Hildebrandt, M., & Jaquet-Chiffelle, D.-O. (2010). Bridging the accountability gap: Rights for new entities in the information society. *Minnesota Journal of Law, Science & Technology*, *11*, 497.

Marcus, G., & Davis, E. (2019). *Rebooting AI: building artificial intelligence we can trust*. Pantheon.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. doi:10.1007/s10676-004-3422-1

Metz, C. (2016). The Sadness and Beauty of Watching Google's AI Play Go. *Wired*. Retrieved from https://www.wired.com/2016/03/sadness-beauty-watching-googles-ai-play-go/

Ministry of Local Government and Modernisation. (2020). *The National Strategy for Artificial Intelligence*. Retrieved from https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=4

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. doi:10.1007/s11948-017-9943-x PMID:28721641

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information Communication and Society*, *23*(5), 1–17. doi:10.1080/1369118X.2020.1713842

Pettit, P. (2001). *A theory of freedom: from the psychology to the politics of agency*. Oxford University Press on Demand.

Pitkin, H. F. (1967). *The concept of representation*. University of California Press. doi:10.1525/9780520340503

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, P. et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3351095.3372873

Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy*, *35*(4), 705–717. doi:10.1111/japp.12274

Sacksteder, W. (1984). Man the Artificer: Notes on Animals, Humans and Machines in Hobbes. *The Southern Journal of Philosophy*, *22*(1), 105–121. doi:10.1111/j.2041-6962.1984.tb00328.x

Sætra, H. S. (2019). Explaining Social Phenomena: Emergence and Levels of Explanation. In Social Philosophy of Science for the Social Sciences (pp. 169-185). Springer.

Skinner, Q. (2005). Hobbes on representation. *European Journal of Philosophy*, *13*(2), 155–184. doi:10.1111/j.0966-8373.2005.00226.x

Solum, L. B. (1991). Legal personhood for artificial intelligences. *North Carolina Law Review*, *70*, 1231.

The Norwegian Data Protection Authority. (2020). *Starter regulatorisk sandkasse for utvikling av ansvarlig kunstig intelligens*. Retrieved from https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-2020/regulatorisk-sandkasse-for-utvikling-av-ansvarlig-kunstig-intelligens/

Tigard, D. W. (2020). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, 1–19.

Wellman, C. H. (2012). Responsibility: Personal, Collective, Corporate. In *A Companion to Contemporary Political Philosophy* (pp. 736–744). Blackwell.

Winfield, A. F., & Jirotka, M. (2017). *The case for an ethical black box.* Paper presented at the Annual Conference Towards Autonomous Robotic Systems. doi:10.1007/978-3-319-64107-2_21

ENDNOTES

1 Also asked by Gunkel (2017).

2 The code could also be written by a machine, which in turn was created by human endeavour. Matthias (2004) also considers machine-written code, but I consider such code also as ultimately attributable to human beings.

Henrik Skaug Sætra is a political scientist working at Østfold University College. He specialises in political theory, and has worked extensively on game theory, environmental ethics. He is currently involved in several projects dealing with the social and philosophical implications of how we employ big data and artificial intelligence in today's society.