When HAL Kills, Stop Asking Who's to Blame

Autonomous machines and responsibility

By Minao Kukita

'Responsibility' is a concept of central importance in ethics. In ethics, responsibility has traditionally been applied only to a human or a group of humans. However, the recent development of ICT has revealed a limitation of this old conception of responsibility. In this paper, we will see what conditions are traditionally required for responsibility attribution, and how these conditions are faced with difficulties by technological developments today, especially those that produce autonomous agents like AIs or robots. Then, we will consider how the concept of responsibility has to change in a society where autonomous machines and humans coexist.

Keywords: autonomous machines, responsibility
Categories: *Automation.*
Corresponding Author: *Minao Kukita*
Email: *minao.kukita@is.nagoya-u.ac.jp*

Introduction

Daniel C. Dennett once asked, 'When HAL kills, who's to blame?' He suggested that it is possible to blame an artificial intelligent system with higher-order intentionality (such as HAL 9000 in the film *2001: A Space Odyssey*), which is the ability to reflect on, think about, or have desires concerning its own mental state.[1] For example, if a machine intended to be a being which desires to be kind to others, the machine could have a kind of higher-order intentionality. While Dennett attempted to explore the theoretical possibility that an artificial intelligent system can qualify as a responsible agent, today this question has gained practical importance. This is not because artificial autonomous systems have acquired a certain level of higher-order intentionality to the extent that we may regard them as praiseworthy or blameworthy, which they do not seem to have acquired yet, but because they are likely to kill.

Car manufacturers and ICT companies across the world are now competing to develop self-driving systems, with Ford recently announcing that its fully autonomous cars with no steering wheels or gas pedals will be in mass production within five years.[2] While the United Nations Convention on Certain Conventional Weapons has debated lethal

[1] When HAL Kills, Who's to Blame? Computer Ethics. Daniel C. Dennett. *HAL's Legacy: 2001's Computer as Dream and Reality*. D. G. Stork (ed.). MIT Press, 351-365, 1997.
[2] CNBC, http://www.cnbc.com/2017/01/09/ford-aims-for-self-driving-car-with-no-gas-pedal-no-steering-wheel-in-5-years-ceo-says.html, Accessed 01/02/2017.

autonomous weapons systems for several years,[3] Israel Aerospace Industries recently disclosed its semi-autonomous uninhabited vehicle for military use called 'RoBattle'.[4] Both self-driving cars and autonomous weapons systems are sure to cause serious damage (including death) to those who are not engaged with them or who are not supposed to be affected by them. One important question in deploying autonomous systems in open situations where they interact with an indefinite number of people is who will be held responsible for the behaviours of an autonomous system, especially when those behaviours lead to unexpected damage. Car accidents are inevitable, and the possibility that non-combatants are unintentionally killed in warfare cannot be eliminated. Deploying autonomous systems in transportation or warfare will make it difficult to identify the person who is responsible for the damage.

In ethics, responsibility traditionally has been applied only to a human or a group of humans. It has been thought that only agents can be held responsible if they are capable of predicting the consequences of actions and intentional decision making. Because only human beings are capable of these things, or at least so it has seemed, it has been argued that only human beings can be held responsible. However, the recent development of ICT has revealed a limitation of this old conception of responsibility. Whether we are conscious of it or not, our recognition, decision-making, and actions are becoming increasingly supported and influenced by technological artefacts. This situation makes it more and more difficult to identify who (or what) is really responsible for the consequences of one's action. For example, are you responsible for overlooking an important message from your colleague which your e-mail management software filtered and classified as a spam? Do you blame yourself or Amazon when you buy a book which Amazon recommended to you, and it turns out to be terrible? Some ethicists are paying attention to the reduced sense of responsibility in a society where our actions are mediated by computers or other complex artefacts.[5] The difficulty becomes more salient as technological artefacts acquire greater complexity and autonomy.

Self-driving cars can be a great benefit to society because they are likely to reduce the number of accidents and energy consumption, mitigate traffic jams, and enable those who cannot drive to use cars. However, the concern about reduced or lost responsibility might hinder the implementation of self-driving systems in a way that allows us to fully enjoy their potential benefits. Therefore, it will be valuable to consider what will or should become of our concept of or practice concerning responsibility in the age of autonomous machines.

In this article, we will first examine our traditional conception of responsibility, and how the traditional conception of responsibility, conditions for attributing it, and practice concerning it are confronted with difficulties based on emerging technologies that produce

---

[3] The United Nations Office at Geneva, http://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument, Accessed 01/02/2017.

[4] Israel Aerospace Industries, http://www.iai.co.il/2013/32981-47061-en/MediaRoom.aspx, Accessed 01/02/2017.

[5] Accountability in a computerized society. Helen Nissenbaum. *Science and Engineering Ethics*, 2(1), 25-42, 1996.

autonomous agents used in open situations where they interact with other artefacts or human beings. Then, we will consider how the concept of responsibility must be altered in a society where autonomous machines and humans coexist. Here we will argue that it is sometimes useless or even harmful to search for someone to blame.

Our focus will be on the original role or function for which the conception of responsibility has evolved. We assume, along with Joshua Greene, that our morality has evolved because it facilitated cooperative behaviours in our ancestors and thereby increased the chance of their community's survival.[6] If so, the same will hold true for our concept of responsibility because responsibility is a central concept in ethics. With the concept of responsibility, we encourage each other to do good and discourage each other from doing harm to others. This is why we have developed and maintained our conception of responsibility. However, emerging technologies make it difficult for the conception of responsibility to fulfil its original function. Therefore, if responsibility is to continue to perform the same job rather than remain the same, we must revise it.

We will suggest that it would be not only useless, but also costly to search for individuals to blame when an accident happens due mainly to the actions of a complex artificial autonomous system or the interactions among such systems. Although humans are not responsible for it, blaming machines serves no purpose because blame only makes sense if the blamed feels guilty and, in response, changes his or her tendencies toward the actions. Instead, we should think more about holding manufacturers or even society responsible to compensate for damages and to improve the systems in order to prevent future harmful events.

Maybe it is hard not to blame someone when something goes wrong. Our emotions are wired to search for the culprit of a mishap and blame and punish him or her. It is helpful to recognise that although this disposition may have been advantageous in the past, when no artificial agents with high autonomy existed, it may not continue to be so in the future.

Finally, we will mention the responsibility gap which is allegedly created by autonomous weapons systems. One of the reasons people are opposed to them is that we cannot identify who is responsible for the war crimes they commit. Although our suggestion might seem to diminish the force of this objection, we will claim that our proposal is not applicable to autonomous weapons.

A standard conception of responsibility and its difficulty

Traditionally, it has been thought that only humans or groups of humans can be held responsible. We do not usually attribute responsibility to animals (though, as mentioned below, there were times when Europeans put animals on trial and punished them). Neither do we hold infants or those with severe mental illnesses responsible for their actions. This

[6] Joshua Greene wrote, 'Morality is a set of psychological adaptations that allow otherwise selfish individuals to reap the benefits of cooperation.' *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Joshua Greene. Penguin Press, p. 23, 2014.

is because responsibility is thought to be limited to those who have certain cognitive and behavioural capabilities, as well as a certain amount of autonomy. One will not be regarded as responsible for anything unless one is able to forecast with a certain precision what will happen as a result of one's actions, make a decision autonomously based on the forecast, and intentionally take action. Thus, we do not hold responsible infants or those with severe mental illnesses, because they do not have the cognitive or behavioural capacities that are needed for one to be responsible. More or less along these lines, philosophers and ethicists have tried to establish the conditions on which one is judged to be responsible.

However, this and similar accounts are problematic. Modern psychology and neuroscience have been revealing that our decision making is largely determined by external factors and unconscious processes that are beyond the agent's conscious or intentional control. Sigmund Freud suggested that most of our behaviours are not the result of conscious contemplation, but of unconscious (often suppressed) desires or impulses. The 'reason' for an action is in fact what the agent makes up afterward in order to rationalise it. Many experiments have been conducted to confirm Freud's theory of unconsciousness and rationalisation. Stanley Milgram showed that ordinary people will do incredibly cruel and harmful things to others when an authoritative person orders them to. Milgram's experiment profoundly shakes our confidence that we will adhere to our conscience when faced with grave immorality. On the contrary, even a tiny excuse could mute our conscience. Benjamin Libet showed that when someone moves her body, the readiness potential[7] precedes one's conscious decision by .35 seconds, suggesting that our conscious decision making has no effect on our actions. According to these studies, against the modern image of an autonomous person who rationally and consciously decides what to do according to his sense of situations and social norms, we do not (or cannot) exert our free will or reason most of the time.

Accumulated scientific findings seem to show that, contrary to the traditional image, we do not have much autonomy and freedom. Some people even doubt whether we have any autonomy or freedom at all.[8] Everything we do is the result of external factors. Every action we take is completely determined by the state of the world a moment earlier. This doctrine is called 'determinism'. One of the possible implications of determinism is that we must be exempted from any responsibility for what we have done. Therefore, pessimistic determinists claim that it is pointless and even wrong to blame someone, for she is not responsible for whatever she did. However, blaming someone seems so natural that it would be difficult for most of us to accept this conclusion.

Faced with this difficulty, a quite different approach to responsibility was proposed by Peter F. Strawson.[9] Strawson objected in his seminal paper to trying to establish rational conditions or theories based on which we can judge whether one is responsible. He contended that when we hold someone responsible, it is not the result of rational judgement based on some external criteria, but of our reactive attitudes, such as resentment, etc., to

[7] The readiness potential is a measure of electronic activity in the motor cortex preceding and leading to muscle movement.
[8] See, for example, S. Smilaminsky, *Free Will and Illusion*, Oxford University Press, 2000.
[9] Freedom and resentment. Peter F. Strawson. *Proceedings of the British Academy*, 48, 1-25, 1965.

her behaviour or intention expressed by it. Such reactive attitudes are natural manifestations of the interpersonal nature of a form of our lives. Social or interpersonal relationships are the network of our reactive attitudes to one another. Therefore, the attempt to discover the conditions on which we can rationally hold someone responsible is misguided, and thus we have no need to give up our usual practices concerning responsibility in the face of the psychological or neuroscientific findings that may subject our autonomy and responsibility to considerable doubt.

More recently, Toshiaki Kosakai, a psychologist at Université de Paris, writes in his book[10] that we do not hold someone responsible for a mishap because she acted on her free will and caused it. Kosakai turns the picture around and asserts that, to the contrary, we regard someone as having free will because we have to hold her responsible. Kosakai clarifies how much our perceptions, judgements, decisions, and actions are subject to external factors. Referring to plenty of examples such as Milgram's experiment, death penalty in Japan and the U.S. and the Holocaust, Kosakai shows how easily ordinary people are brought to do incredibly cruel or irrational acts depending on situations, and explains the psychological mechanisms and social structures behind them. Then, Kosakai concludes that autonomous decision making is an illusion, and thus that responsibility is a fiction. Kosakai holds that explaining responsibility in terms of a causal relationship (i.e. one is responsible for something if she caused it) is wrong.

However, being a fiction does not imply that it is dispensable. Kosakai says that our society is supported by many fictional stories, such as about money or nation.[11] Responsibility is one of them. Getting rid of some of them may not be without considerable cost. Thus, Kosakai does not propose to improve our practices concerning responsibility despite it being a fiction. We, as humans, cannot help thinking about our actions within the framework of autonomy, causality, and responsibility. Faced with a serious mishap, according to Kosakai, we search for the culprit, and try to blame and punish her in order to counterbalance the damage. It is as if the person who is blamed and punished were a scapegoat dedicated for the order of society.

Both Strawson and Kosakai place more priority on practice than on theoretical explanations. They both think that it is hard (or maybe costly) to improve drastically our practice concerning responsibility. Admitting the difficulty, though, we shall consider the possibility of improvement, and its potential benefits or other consequences.

How blaming works

Strawson pointed out that reactive attitudes are accompanied by certain emotions such as resentment. However, he did not ask further how we have come to have such emotions. Kosakai analyses psychological and sociological mechanisms behind our practices

[10] *Responsibility as a Fiction*. Toshiaki Kosakai. Tokyo Daigaku Shuppankai, p. 151, 2008. (In Japanese, the original title is 『自由という虚構』)

[11] For the power and effectiveness of fictions, see Yuval Noah Harari, *Sapiens: A Brief History of Humankind*, Harper, 2015, especially, chapter 2.

concerning responsibility. However, Kosakai's focus is on the modern conception of responsibility, which is closely connected with the modern notion of an 'autonomous individual'. It is contrived to help to control and maintain the order of society. As for pre-modern societies, Kosakai thinks that religion and superhuman beings played the same role. However, the practice of blaming and praising individual people surely existed before the modern notions of autonomy and responsibility were born. We will consider this more primitive practice of blaming and praising, as well as accompanying emotions.

The notion of responsibility is now embedded in our practice in which someone predicts consequences of her actions, makes decisions, and controls the course of her actions so that desired results will be brought about or undesired results will not, and if her actions have led to some harmful consequences, the community may condemn and punish that person. However, prior to the modern notion of responsibility based on individual rationality, free will, and autonomy was established, we had similar practices—practices in which we blamed, praised, punished, rewarded, or forgave someone for doing something bad or good. Why have humans developed such a routine of blaming or praising?

It does not seem to be limited to certain cultures and is fairly universal. Precursory tendencies of this practice can be found even in preverbal children and some species of primates. In one study, preverbal children were shown animation movies in which simple geometrical figures interacted with one another.[12] Some attacked others, while others prevented the attack and saved the one being attacked. After watching the videos, the children developed a preference for those who saved others under attack over those who did not. In another study, common marmosets were shown interactions between two human actors.[13] In one scenario, an actor offered a gift to the other, and then the latter offered a gift in return; in the other scenario, one offered a gift but the other do not reciprocate. After watching the interactions, marmosets were offered food from these actors. The marmosets accepted food less often from the non-reciprocators than the reciprocators. Common marmosets have relatively small brains compared with other primates such as chimpanzees or gorillas and are considered to be less intelligent, but they are known to be more pro-social. The authors suggest that it is not general intelligence or higher cognitive ability but the cooperative and pro-social tendencies that enable them to evaluate the interaction between others and identify who is cooperative and who is not.

We seem to have a natural tendency to like and praise people who are cooperative or altruistic and to dislike people who are selfish or aggressive. Following Joshua Greene's idea that morality developed because it promoted cooperative behaviours within groups and survival values, let us assume that these tendencies have also contributed to our cooperative behaviours. It also seems natural to assume that these tendencies developed into our more sophisticated practice of blaming and praising. With the practice of blaming and praising, people are encouraged to take action that is beneficial to the community, as well as discouraged from taking action that is harmful to others. The role of blaming and

[12] Preverbal infants affirm third-party interventions that protect victims from aggressors. Yasuhiro Kanakogi, Yasuyuki Inoue, Goh Matsuda, et al. *Nature Human Behaviour*, 1(0037), 2017.
[13] Marmoset monkeys evaluate third-party reciprocity. Nobuyuki Kawai, Miyuki Yasue, Taku Banno, et al. *Biology Letters*, 10(5), 2014.

praising is to direct people toward cooperative behaviours that promote total utility within the community. It is because of this function that the practice developed in human society. Later, this practice combined with the modern notion of individuals who have rationality, freedom of action, and autonomy.

If we put more emphasis on the original function of blaming and praising and the origin of responsibility rather than what we are thinking, feeling, or doing with this concept in practice, and if the traditional understanding of responsibility has become obsolete and cannot fulfil its function well in present situations, we need to revise the concept. Perhaps, progress in science and technology now calls for a change in our understanding of responsibility.

A need for revision of the concept of responsibility

As we saw in the previous section, one key component of responsibility is causality: causal contribution is required for one to be held responsible. However, it is now becoming harder and harder to attribute the responsibility for some serious incidents to certain autonomous individuals or groups that can be identified as main factors causing them. There are two main reasons for this. First, as we saw in the previous section, psychology and neuroscience have cast doubt on how autonomous individual humans are, i.e. how much freedom of action we have. Some even believe that we have no free will at all. Second, the role of technology in our decision-making and actions is becoming greater. Here, we will focus on the second issue. Even if we admit individual autonomy and free will, the technological developments today are making it harder to identify who is responsible.

Since ancient times, we have always been supported by tools when we perceive, think about, navigate in, and act upon our surroundings. Since the industrial revolution, the extent to which we are supported by tools has been enlarged at an ever-increasing rate. I am hardly able to do anything important without the aid of technological artefacts. Yet, thus far, we have taken it for granted that technological artefacts are just instruments, and that the user is held responsible for anything that results from the use of such instruments. However, the recent development of information technology—robotics and artificial intelligence in particular—is reaching a critical point where we may or should change such a common understanding.

Technological artefacts are media between the user and object in the surrounding world. They convey information to the user and transform the objects in the surrounding world according to the user's intention. Following Luciano Floridi, a 'prompter' is an object on which the user acts on through the mediating technology.[14] When the user is a human being and the prompter is something in the natural world, Luciano Floridi calls the technologies between them 'first-order technologies'. If you use an axe to cut down a tree, the axe is a first-order technology. When the user is a human being and the prompter is a technological artefact, the technologies between them are called 'second-order technologies'. When you hit a nail with a hammer, the prompter is the nail, and the hammer is a second-order

[14] *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*. Luciano Floridi. Oxford University Press, p. 25, 2014.

technology. When both the user and the prompter are technological artefacts, the technologies between them are called 'third-order technologies'. In third-order technologies, there are no humans involved.

Take, for example, the relationship between your personal computer, your router, and the server. In this case, your personal computer can be viewed as the user, the server the prompter, and your router the third-order technology between them. It is true that, if you view this situation from another perspective, you are the user of the system composed of your computer and the router, and the prompter is the server. Then, the computer and the router together constitute a second-order technology between you and the server. In this way, perhaps, every technology can in principle be embedded in a scheme of the first- or second-order technology.[15] Still, it is important to pay attention to what kinds and amount of interactions are being made in third-order technology. For it is exactly the kind and amount of such interactions that are making it difficult to attribute responsibility—to identify who is responsible for what happened. This is because a third-order technology usually works without direct human supervision or control. It works according to its design and increasingly interacts with other technologies and human beings.

In this regard, robots, artificial intelligent systems, and IoT technologies demand special attention. Artificial intelligent systems are becoming smarter, more autonomous, and able to perform cognitive and intelligent tasks that we previously thought only humans could perform. Robotics can embody artificial intelligence in the real physical world. IoT technologies add complexity to the whole network of intelligent agents. Being complex means that their behaviours are difficult (and sometimes impossible) to predict. In addition, sophisticated interfaces between technologies and humans ingeniously hide what is going on inside, and enable us to reap the harvest of complex technologies without truly understanding what they are doing. In short, technologies today can perform tasks that only humans could perform previously, interact with one another without our control or our knowledge of what they are doing, and work in such a complex way that nobody can exactly predict what will happen.

These technologies are sources of great convenience, power, and wealth, and we will increasingly make use of and become dependent upon them in more and more situations. However, there are several obvious downsides. One is safety. Recall the 2010 flash crash, a sudden stock market crash in the U.S. One of its factors was high-frequency trading algorithms employed by many stock trading firms. Complex and fast interactions among machine traders contributed to the possibly catastrophic stock market crash. Fortunately, the average stock price recovered after 36 minutes. However, this incident is a clear indication of the possibility that a devastating effect can follow the advance of artificial intelligence. Another notable incident was the fatal car crash involving a self-driving car. Last year, a Tesla Model S, with its 'Autopilot' (driver-assistance system) on, crashed into a tractor trailer, causing the death of the driver of the former. It was said that, while the tractor trailer was crossing the road in front of the Tesla, the driver-assistance system mistook the trailer for a road sign, or failed to recognise it against the background of the

[15] It is possible that the human user can be absent temporarily. Imagine that the user suddenly dies from a heart attack, and her personal computer continues the transaction with the server for a while.

sky. Federal auto-safety regulators carried out an eight-month investigation and found no defects in the autopilot system of the car. They warned that driver-assistance systems such as the one employed by Tesla Model S can be relied on to react to what happens on the road only in some situations, and the officials said that automakers need to be clear about how their systems should be used.[16]

If we are to make full use of artificial intelligent systems, we have to face the possibility of serious accidents. An important question is this: Who is responsible for a deadly accident?

If the user was operating the system properly, an obvious answer to this question is to distribute responsibility between the designers and manufacturers. However, a technological artefact today can be a complex assembly of many component technologies, and it functions as a result of sophisticated interactions between its parts. Moreover, today's artificial intelligent systems often use some kind of machine learning techniques, and their behaviours are dependent not only on their design but also their learning. Recall Tay. It was a chatbot on twitter, developed by Microsoft. It learned what to say from other twitter users, and in less than 20 hours since its first tweet 'hellooooooo world!!!'[17], Tay began to utter racist sentences. Who is responsible for the nasty and offensive things Tay uttered? Programmers, Microsoft, Twitter, users who talked with Tay, or Tay itself?

It is not easy to determine how to distribute responsibility in such a case. It is impossible for designers to predict every situation and way in which users will deploy their artefacts. As a result, designers or manufacturers cannot possibly predict how their products will behave in every practical situation. Therefore, it is also difficult to attribute responsibility, conceived in traditional terms, to the designers, manufacturers, and so on. The same holds true of other parties.

Some ethicists insist that we should update concepts like 'responsibility' or 'agency' so that they will be effective in today's world. Deborah G. Johnson, for example, argues that we cannot dismiss the role played by computational artefacts as irrelevant.[18] Although artefacts do not have intention, the intentions of designers are inscribed in them, which human users activate. Luciano Floridi and J. W. Sanders propose that we should acknowledge agency to those artefacts that have enough autonomy and adaptability, and interact with other agents.[19]

Nevertheless, such ethicists do not think that we can attribute responsibility to any artefacts (at least until they have their own intentions, according to Johnson). While Floridi and Sanders claim that we can hold some artefacts accountable, in the sense that they are one of the main factors contributing to some event, they do not think that we can hold them

---

[16] The New York Times, https://www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html?_r=0, Accessed 10/04/2017.
[17] Actually, 'o' in 'world' is the emoji of the earth.
[18] Computer Systems: Moral Entities but not Moral Agents. Deborah G. Johnson. *Ethics and Information Technology*, 8: 195-204, 2006.
[19] On the Morality of Artificial Agents. Luciano Floridi and J. W. Sanders. *Minds and Machines*, 14(3): 349-379, 2004.

responsible in the sense that they are to blame. This is reasonable. Blaming someone is effective because it might mean punishment for him or her, and punishment is effective because people have self-interests. However, no artefact that currently exists has any self-interest.

It is true that some researchers consider the possibility that a robot could be a bearer of right and/or responsibility. For example, the European Parliament is considering giving a form of 'electronic personhood' to the most capable robots and artificial intelligence, in order to ensure rights and responsibilities for them. Jerry Kaplan also claims that robots in the future could be treated as having responsibility of their own, though he admits that, for the time being, only humans are responsible.[20] Citing a case where the U.S. Department of Justice filed not only a civil case but also a criminal case against BP, an oil company that caused a massive oil spill in the Gulf of Mexico (called 'Deepwater Horizon oil spill') in 2010, Kaplan claims that moral agency does not require consciousness or feeling. A robot can be a moral agent if it is 'capable of recognizing the moral consequences of its actions' and 'able to act independently'. Kaplan even holds that inflicting punishment on robots will be effective, just as it is effective for corporations. Humans, corporations, and robots share a common characteristic: they all have a purpose or goals. Thus, he argues, if a robot is disabled from pursuing its purpose (for example, by deleting its data or learned 'knowledge'), it will amount to punishment in effect. This treatment may be useful because the existence of such punishment will send a message to the designers, trainers, and users that it makes sense to ensure that their robots are beneficial and not harmful to others, and will lead to improvement of the robots' behaviours.

Kaplan's proposal is particularly interesting because it does not completely identify robots as corporations. If robots are to be treated exactly in the same way as corporations are (i.e. as having legal personhood), they can have rights and responsibilities right now, without acquiring the capability of 'recognizing the moral consequences of [their] actions' or the ability to 'act independently'. After all, we do not think that corporations have either of them. It would also be unnecessary to inflict punishment on them other than the imposition of fines. As Masahiro Kobayashi, a lawyer at Hanamizuki Law Office in Osaka, Japan, once pointed out in a lecture,[21] giving legal personhood to robots will mean no more than establishing an appropriate insurance that will cover future damages caused by the robots. It seems that Kaplan (consciously or not) regards artificial intelligence and robots as different from either human beings or corporations, as something in between. Accordingly, Kaplan's conception of responsibility is different from the existing ones. Kaplan's confusing account of responsibility suggests that we need to reconsider the notion of responsibility in this age of robots and artificial intelligence.

[20] *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. Jerry Kaplan. Yale University Press, Chapter 2, 2015

[21] At a forum on artificial intelligence and self-driving cars held jointly by Nagoya University and DENSO, a Japanese car-component manufacturer, 4 December 2016, in Kariya, Japan.

Rethinking responsibility

Though Kaplan uses the word 'purpose' or 'goal' for robots, of course, robots have neither purpose nor goal of their own. Their purposes or goals are provided by the human beings who design or utilize them. So, if punishments are inflicted on robots, they are not actually inflicted on robots, but rather on the human beings involved with them. For current complex artificial intelligence and robots, there is a large number of people involved with them. Thus, any punishment will be divided into tiny pieces and distributed among those people. The problem is that punishment has a threshold under which it is hardly effective. That is, a one dollar fine for one million people does not have the same effect as a one million dollar fine for a single person. Therefore, it is uncertain that Kaplan's proposal that robots should be responsible is effective.

Here, we will consider an alternative approach to the problem of responsibility in the age of autonomous machines. Let us state the problem concisely. Today, we are surrounded by technological artefacts that are functioning autonomously—without explicit control and/or supervision by human operators—and the number and variety of such robots are increasing at a surprising rate. In addition, unlike traditional industrial robots, they are not confined within factories where they interact only with factory workers and other factory robots. Today's robots are functioning in unlimited (real or virtual) open spaces where they interact with indeterminate people and other machines. Moreover, some of them can learn from their past experiences and adjust their behaviours. As robots' autonomy, complexity of interactions, and adaptability increase, it is becoming more and more difficult for anyone to predict their behaviours, and, as we have seen in the instance of the 2010 flash crash in the U.S. stock market, the fatal car crash of a Tesla Model S, and the racist utterances by the chatbot Tay, autonomous machines can have harmful effects. Then the problem is who is responsible for such damage caused by autonomous machines. To cite Daniel C. Dennett, 'When HAL Kills, Who's to Blame?'

We cannot answer the question by analysing the notion of responsibility, for we are now living in an environment that is quite different from those in which the notion of responsibility was born, developed, and established. Our practice of blaming and the notion of responsibility or guilt are not fixed. They have been changing as our societies, civilisation, and understanding of the world and human beings develop. For example, in the Middle Ages, European put animals on trial and punished them. Today, most societies have no such practices or conception of responsibility. In the new environment, the traditional notion is confronted with difficulties. Therefore, we have to reconsider the way responsibility ought to be, taking into account its original function and current situations.

We assumed above that blaming and praising are more primitive than the modern notion of responsibility, and hence our practice concerning blaming and praising has the function of encouraging members of society to take beneficial actions and discouraging harmful ones. However, the conventional practice does not fulfil the function well in current situations where human decision-making and actions depend heavily on technologies, and will fulfil it even less in the future. This is because it is possible that serious harm may occur as a result of using a technological artefact, and there will be no one (or at least no human) that can be identified as the culprit, and it makes no sense, at least for now, to

attribute responsibility to machines. To make responsibility function properly, we need a new way of thinking about responsibility. However, before we do so, two issues must be addressed.

First, we will distinguish two kinds of responsibilities: the forward responsibility and the backward responsibility. The forward responsibility is one's duty or obligation to do (or to refrain from doing) something. The backward responsibility is concerned with one's actions in the past. One is usually thought to be responsible for things that are seemingly caused by her conscious and intentional acts. In our practice, these two responsibilities are closely connected with each other. If one is backwardly responsible for something, then she will bear forward responsibilities to account for those past acts and/or to compensate for it. Conversely, if one failed to fulfil a forward responsibility, she will bear a backward responsibility for that failure. It is the backward responsibility that is associated with blame and punishment, and we are mainly concerned with it, when we propose that we should improve our practice concerning responsibility. In other words, we will leave the concept of forward responsibility almost intact.

Second, we will also distinguish four types of disasters. We usually distinguish two kinds: natural and man-made disasters. As for a natural disaster, such as an earthquake or a lightning strike, we do not usually blame anyone for it, because we have no way to exactly predict or prevent it. If a disaster is man-made, we blame those who contributed to it. However, in a society immersed in advanced science and technology, the line between these classes of disasters is being blurred. This blurring comes in two ways. On the one hand, the advance in science has enabled us to predict or control some natural disasters that had been formidable for humans. As a result, when such a disaster causes serious damage, it is likely that someone will be blamed for the failure in predicting or controlling it. We will call this type of disaster 'semi-natural'. For example, six Italian scientists who had reassured the public before the deadly earthquake in L'Aquila were convicted of manslaughter, though they were later exonerated. There had been 'swarm' tremors before the earthquake, and people were worried. Their reassurance might have led some of the victims to stay indoors rather than go to a shelter on the night of the earthquake, thus causing their deaths. On the other hand, technologies and social systems have become so huge and complicated that even experts can no longer adequately understand or control them. In such complex systems, a catastrophic disaster will result from an accumulation of tiny errors or irregularities, each of which is insufficient to cause the disaster. Charles B. Perrow, a sociologist and emeritus at Yale University, calls such disasters 'normal accidents'. The word 'normal' here means that such accidents inevitably result from usual operations. We will also call this type of disaster 'normal'. Thus, we distinguish four types of disasters: natural, semi-natural, normal, and man-made. Here we will focus on the third class of disasters. Thus, we have in mind such complex and huge-scale technologies that their implementation into society will have a great impact on a wide range of people.

The increase in both the autonomy of machines and the interconnectedness of distributed systems is today's technological trend, but both tend to add to the complexity of existing systems, the uncertainty of their behaviours, and the range and magnitude of disasters caused by the systems, as we saw in the 2010 flash crash—a perfect example of a normal

accident. However, this technological trend seems inevitable,[22] and thus we have to learn to live with these trends. For this purpose, we hold that rethinking responsibility is helpful.

Now, our proposal of rethinking responsibility is as follows:

A. Responsibility shift from individuals to society

Emphasis should be placed on responsibility for social decision-making concerning the technology in question, rather than on responsibility for individual incidents caused by using the technology. When we consider introducing a technology into society, we have to ask and discuss what influences it will have on society and what risks there will be. In addition to conducting ordinary cost-benefit analysis, we have to question whether it is done in a morally permissible way, rather than in purely economic terms. Ethical and democratic risk analysis are required to reflect social values and should involve representatives from a variety of stakeholder classes influenced by the introduction of the technology. Moreover, the details of the risk analysis must be public.

Of course, if the harm overrides the benefit, the technology should not be introduced. After introduction, risk analysis should be continually conducted with newly obtained information, and if unexpected costs arise and the total benefit is overridden, some measures, including total abolition, should be taken. In addition, it is desirable that its introduction would not inflict damage on people who would be safe without the technology. At least, imposing risks on those people who have no chance of receiving the benefit of the technology is unfair and unethical.

B. Stop asking who is to blame, and focus on why it happened

If the technology is introduced after going through the process of social decision making as mentioned above, and an accident happens without the fault of anyone, we should not waste too much time discussing who is to blame. Rather, we should try to understand why undesired results happen and improve the system. For this purpose, the system should be transparent. The burden of compensating for the harm should be imposed on the beneficiaries, according to how much they benefit from the technology. Beneficiaries include not only those who are paid for developing, administrating, maintaining, or selling the technology, but also society as a whole, which can reduce costs by using the technology. For example, if the self-driving system mitigates the number of car accidents, traffic jams, energy consumption, and environmental damage, society as a whole is considered to be a beneficiary.

The designers or administrators should try their best to unravel and explain why the trouble has occurred. Because it may have resulted from an unusual situation in which the system was deployed, the user may also be held partly accountable. However, we need or should not blame anyone. It may be a natural tendency (or second nature) to find someone to blame when something bad happens. Although such tendency may be good in that it helps people improve their behaviours to refrain from taking potentially harmful actions, it is worth

[22] These are two of Kevin Kelly's inevitable technological trends: 'cognifying' and 'accessing'. Cf. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future*. Kevin Kelly. Viking, chapters 2 and 5, 2016.

questioning whether that tendency is beneficial or appropriate in every situation. Blaming or punishing will likely be useless because no particular person's fault brought about the accident, and thus it will not lead to any improvement of the system. Rather, blaming discourages people from cooperating in the efforts to unravel why the accident happened.

In fact, our proposal is not limited to autonomous machines, but can be applied to complex and highly influential technologies in general, and even to complex social systems. In this regard, autonomous machines are not so unique. It is not unusual for a man-made system not to function well due to no one's fault. If so, arguments for attributing responsibility to robots may be a result of unnecessary anthropomorphism.

The 'no-blame policy' will be confronted with criticism that it is problematic to hide the source of responsibility. Our proposal is not incompatible with such criticism. We, too, admit that we should make it clear in advance where the buck stops as far as possible. We only contend that it is impossible to preclude the possibility that there is no one who is to blame, though it is desirable that such a situation not arise. Just like pursuing 'zero risk', trying to always put someone in the responsible position can be rather costly and harmful.

Relevance of our proposal to autonomous weapons systems

One of the most threatening applications of artificial intelligence is autonomous weapons. This issue may appear to be relevant to our current discussion at first sight because responsibility is one of the focal points in the discussion of autonomous weapons systems.

Some people object to autonomous weapons systems because they are likely to create the 'responsibility gap'. Namely, when these weapons mistakenly attack non-combatants, it will be difficult to decide who is responsible for the attack.[23] Other people are opposed to the responsibility-gap argument, insisting that, even if behaviours of autonomous weapons systems such as highly automated drones are difficult to predict or control, the programmers should recognise the limitation and risk of the systems and should explain them to those who deploy or operate them. The commander and operator should take into account the limitations and risks of deploying such systems and be prepared for the unexpected.[24] Thus, they argue, the programmers, commanders, and operators should be responsible, and there is, therefore, no responsibility gap here.

The no-blame policy may be criticised by both sides. On the one hand, the proponents of the responsibility-gap argument may think that the no-blame policy instructs people not to care about the responsibility gap, and thus weakens their argument against autonomous weapons. For them, the no-blame policy may help to promote autonomous weapons. On the other hand, the opponents of the responsibility-gap argument may think that the no-blame policy is simply wrong, because there must always be someone to blame.

---

[23] Cf. Killer robots. Robert Sparrow. *Journal of Applied Philosophy*, 24 (1), 62–77, 2007.

[24] Cf. Drones, automated targeting and moral responsibility. Alex Leveringhaus. Ezio Di Nucci and Filippo Santoni de Sio (eds.), *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*, Routledge, 169-181, 2016.

Note that our argument for not assessing blame is only applicable to technologies that are intended to contribute to the improvement of average living conditions for people involved with them, and are introduced by democratic decision making after informing people of the benefits and risks. Also, if the technology is designed to inflict harm on those people who have no chance of benefiting from it, its introduction is not justified. However, military technologies are essentially intended to do harm to some people in order to benefit others. Thus, our argument does not address the 'responsibility gap' issue of military technologies. We agree with the proponents of the responsibility-gap argument that, if autonomous weapons lead to the responsibility gap, we should not deploy autonomous weapons. Meanwhile, we also agree with the opponents who argue that we should always hold someone responsible for the deployment of autonomous weapons so that no responsibility gap will be created.

One might say that some weapons are not intended as a means of destruction, but of deterrence. For example, in an application for a patent, Nicola Tesla wrote that 'the greatest value of my invention will result from its effect upon warfare and armaments, for by reason of its certain and unlimited destructiveness it will tend to bring about and maintain permanent peace among nations.'[25] If so, can we possibly apply the no-blame policy to autonomous weapons?

Maybe yes, but only if the deterrent weapon is introduced through a democratic decision-making process involving those people who are subject to its influences. Because literally every country might be affected by the deterrent technology, there must be a global agreement on the introduction of it. If there is such an agreement, the no-blame policy can be applied. It would be no one's fault if World War III broke out and human civilisation ended. However, it is hard to imagine that to be the case. The deterrence theory assumes that each individual state behaves as a logical and selfish decision maker and, in addition, that it knows that every other state does the same. Meanwhile, the no-blame policy is only applicable to a situation where every actor not only considers its own benefit, but also the collective benefit. Accordingly, the deterrence theory and the no-blame policy have essentially incompatible natures. If the world should agree in a democratic and ethical way, it is more likely that it would decide on disarmament than on mutual assured destruction.

Therefore, the no-blame policy has little relevance to the responsibility issue on autonomous weapons.

Conclusion

The modern notion of responsibility is the combination of our natural tendency to search for someone to blame when a bad thing happens and the concept of an autonomous individual who has the freedom of choice. The notion of responsibility is now faced with difficulties. For one thing, recent scientific findings have revealed that we are not as autonomous as we once believed. Additionally, the great extent to which today's complex and large technologies mediate and support our decisions and actions are making it harder to know who is to blame when an accident happens. Autonomous technologies will further

[25] Method of and apparatus for controlling mechanism of moving vessels or vehicles. Nicola Tesla. US Patent 613809, 1898.

complicate the problem; in the age of autonomous machines, there are some situations in which we could blame anyone for an accident.

Our traditional practice of blaming and concept of responsibility will no longer stand. Therefore, in this paper, we proposed a new practice of blaming: the no-blame policy for accidents resulting from the normal use of technologies which are introduced by democratic and ethical decision making. Admitting that we should make clear who is responsible as much as possible, we emphasised that we could not eliminate the possibility that searching for the culprit will do more harm than good.

We also considered the relevance of our proposal to the dispute concerning autonomous weapons, where the 'responsibility gap' is a focal point, and argued that the no-blame policy is not applicable to autonomous weapons because they are essentially intended to do harm to those people who have no chance of gaining any benefit from the technologies.

Acknowledgement

This study is supported by JSPS Grant in Aid JP16H03341 and JP16H03343.