AI, agency and responsibility

Johnson, Deborah; Verdicchio, Mario

Published in:
AI & Society

DOI:
10.1007/s00146-017-0781-9

E-pub ahead of print: 30/09/2019

Document Version
Publisher's PDF, also known as Version of record

AI & SOCIETY
https://doi.org/10.1007/s00146-017-0781-9

OPEN FORUM

AI, agency and responsibility: the VW fraud case and beyond

Deborah G. Johnson[1] · **Mario Verdicchio**[2]

Received: 30 July 2017 / Accepted: 14 December 2017

Abstract
The concept of agency as applied to technological artifacts has become an object of heated debate in the context of AI research because some AI researchers ascribe to programs the type of agency traditionally associated with humans. Confusion about agency is at the root of misconceptions about the possibilities for future AI. We introduce the concept of a triadic agency that includes the causal agency of artifacts and the intentional agency of humans to better describe what happens in AI as it functions in real-world contexts. We use the VW emission fraud case to explain triadic agency since in this case a technological artifact, namely software, was an essential part of the wrongdoing and the software might be said to have agency in the wrongdoing. We then extend the case to include futuristic AI, imagining AI that becomes more and more autonomous.

Keywords Agency · Artificial intelligence · Autonomy · Ethics · Future · Technology

1 Introduction

A good deal of attention is now being given to artificial intelligence (AI) and its potential to be used in so many domains of human life. Although AI has many forms, it is generally correlated with the concept of autonomy, that is, AI systems are most often seen as autonomous to one degree or another (Castelfranchi and Falcone 2003). The term 'autonomy' is used by AI researchers as a metaphor to refer to a variety of different types of computational behaviour (Maes 1990; Pfeifer 1995; Steels 1995; Franklin and Graesser 1997; Dignum 1999; Dormoy and Kornman 1999; Jennings et al. 2000; Lin et al. 2004; Tentori et al. 2006; Colburn and Shute 2008; Stone et al. 2010; Noorman and Johnson 2014), but the multiplicity of its meanings (both for AI researchers and non-experts) can lead to miscommunication: 'autonomy' often suggests to those in the media and the lay public something out of human control, something worthy of concern and even fear (Johnson and Verdicchio 2017).

On some futuristic accounts AI will not just be out of the control of its designers but will become so sophisticated that it will, in some sense, take over the world (Barrat 2013; Dowd 2017). For example, in his book "Superintelligence", Nick Bostrom speculates that AI will one day reach such a high level of development that it will start giving itself new goals and will build new artifacts to reach them, all this independently of humans, who will become irrelevant at best, extinct at worst (Bostrom 2014). Not coincidentally, at the same time that computer scientists are speculating about developments in AI, popular media—sci-fi movies, TV series, short stories and novels—are playing with and portraying AI robots that look and behave like humans (e.g. *Ex Machina, Humans, Westworld*).

The discussion of AI as autonomous is connected to notions of agency since, both in scholarly speculations and in fiction, the pivotal point seems to be when AI artifacts start acting "on their own", that is, when their agency somehow transforms from the causal efficacy of objects to the intentional action that is usually only associated with conscious, intelligent beings (Johnson and Noorman 2014). Again, not coincidentally, although the focus of attention is most often on autonomy, notions of agency are also being discussed, challenged, and changed in academic discourse (Powers 2013; Sayes 2014).

In this paper, we want to focus on agency and we want to bring discussion of agency 'down to earth', so to speak. Firstly, we are concerned not just with agency in some

✉ Mario Verdicchio
mario.verdicchio@uws.ac.uk

[1] Department of Engineering and Society, University of Virginia, Thornton Hall, 351 McCormick Road, Charlottesville, VA 22904, USA

[2] School of Media, Culture and Society, University of the West of Scotland, Ayr Campus, Ayr KA8 0SX, UK

Published online: 11 January 2018

abstract or ontological sense; we are concerned with the notion of agency as it comes into play in issues of responsibility (ethical and legal). To do this we begin with the Volkswagen (VW) emission fraud case (Hotten 2015), which we analyze with an eye to understanding the agency of a computational artifact. We then use this analysis as a basis for assessing speculation about futuristic AI agency. The VW case is embedded in a legal context, but our analysis is concerned with a more general view of responsibility, both ethical and legal.

Agency is traditionally defined as the capability of an entity to act (Schlosser 2015). In the most common accounts of agency, a distinction is made between behavior produced by mental states (intentional action) and behavior that can be explained in terms of material causal relations. It is the distinction, for example, between a person telling another person it is time to go and an alarm clock going off to alert a person that it is time to go. The former kind of agency is traditionally seen as exclusively human, while the latter is attributed to artifacts as well as humans. This distinction is fundamental when it comes to responsibility because human agents can be considered responsible (accountable, duty-bound, praiseworthy/blameworthy) for their actions, both ethically and legally, while artifacts cannot. For example, a person might be praised for remembering that it was time to go and then alerting another, while the alarm clock would only be considered causally responsible for alerting the person; the alarm clock would not be considered morally praiseworthy for doing so, or legally liable for failing to do so (except in a joking way).

The concept of agency as applied to technological artifacts has become an object of heated debate in the context of AI research because some AI researchers ascribe to programs the type of agency traditionally associated with humans (Allen and Wallach 2012; Bostrom 2014; Malle and Scheutz 2015; Omohundro 2016; Yampolskiy 2016). Although these researchers rarely affirm an identity between the agency of programs and the agency of humans, they do not draw any distinction either, except to point out the greater efficiency and efficacy of the former, thanks to its being technological. This is controversial in part at least because of the connection between agency and responsibility. If computational artifacts are agents, it would seem that they could be moral agents and bear responsibility for actions, especially in cases of technological mishap.

In this paper, we argue that confusion about agency is at the root of misconceptions about the possibilities for future AI. To begin our argument, we distinguish two types of agency and show how the discourse on AI slips from one kind to another. We argue that this slip is problematic and leads researchers to predict unrealistic futures. We then introduce a third kind of agency that includes the first two and better describes what happens in AI as it functions in real-world contexts. This third model of agency is a heuristic for tracing responsibility in technological mishaps. We use the VW emission fraud case to explain the three kinds of agency since in this case a technological artifact, namely software, was an essential part of the wrongdoing and the software might be said to have agency in the wrongdoing. We then extend the case to include futuristic AI imagining AI that becomes more and more autonomous.

We think that a better understanding of agency can provide a clearer view of responsibility when humans act with technology, be it artifacts of today or futuristic AI.

2 Types of agency

In the past, the term 'agency' referred to the capacity to act; action was distinguished from mere behavior, as in the case of the behavior of artifacts; and only those entities with the capacity to act were called 'agents'. Agency was presumed to apply exclusively to humans whose acts were seen as resulting from intentions. More recently, however, this notion of agency has been contested. AI researchers have adopted the idea of "artificial agents" (Jennings et al. 1998; Ferber 1999; Weiss 1999; Wooldridge 2002; Floridi and Sanders 2004; Floridi 2008). In AI, the operations of a program or a robot are seen as actions, which means that programs and robots count as agents. They are 'artificial' in that their actions are computational and embodied inside electronic circuits, as opposed to the 'natural' actions performed by humans. What, then, are the implications of saying that AI artifacts are (artificial) agents?

The current discourse outside AI draws on multiple notions of agency. In the *Stanford Encyclopedia of Philosophy*, for example, Schlosser distinguishes a broad and narrow meaning of agency:

> In a very broad sense, agency is virtually everywhere. Whenever entities enter into causal relationships, they can be said to act on each other and interact with each other, bringing about changes in each other. In this very broad sense, it is possible to identify agents and agency, and patients and patiency, virtually everywhere. Usually, though, the term 'agency' is used in a much narrower sense to denote the performance of intentional actions. (Schlosser 2015)

Here Schlosser generously expands the notion of agency to include entities that are causally efficacious. Entities that act with intentions are causally efficacious, though the causal chain that explains their actions initiates from intentions. Later we will discuss claims made by AI researchers as to the intentions of artificial agents.

Literature on agency now spans a number of disciplines including Science, Technology and Society (STS) studies,

where the notion of agency is extended in a way that is parallel to Schlosser's first sense of agency. For example, actor network theory (ANT) introduces a neutral term, 'actants' to include human and non-human things (including nature, ideas, relationships, as well as artifacts) that contribute to the production of a state of affairs (Latour 1996; Law and Hassard 1999; Sayes 2014). Latour is a leading proponent of ANT and as Sayes (2014) explains: "Latour (2005: 71) maintains that one need only ask of an entity '[d]oes it make a difference in the course of some other agent's action or not? Is there some trial that allows someone to detect this difference?' If we can answer yes to these two questions, then we have an actor that is exercising agency—whether this actor is nonhuman or otherwise" (p. 141).

In this paper, we draw on Schlosser's two notions of agency, referring to them as causal agency (broad) and intentional agency (narrow), and we add a new type, which we call triadic agency. Triadic agency is introduced as a way of capturing the type of agency that is at work when humans act with technological artifacts. We use triadic agency as a heuristic device to reveal the role of humans and artifacts in producing events. Although nature (e.g. ecosystems) and ideas (e.g. fairness) shape what happens, our focus is on humans and artifacts. As we will explain later, triadic agency not only expands what can be said about them, it grounds ascriptions of responsibility, both ethical and legal.

2.1 Causal agency

As already explained, causal agency has to do with causality. When artifacts are said to be agents in this sense, the emphasis is on their causal efficacy. This usage draws attention to the important role that artifacts have in the causal chain that produces states of affairs. This type of agency fills a gap since too little attention has been given to the ways in which technology shapes what happens in the world. Technological artifacts can powerfully affect social arrangements, relationships, institutions, and values. Treating artifacts as agents properly frames them as significant constituents of the human world.

Turning now to the VW emission fraud case, we can illustrate causal agency for in this case an artifact, i.e., software, had an essential role in the wrongdoing. In 2015, the US Environmental Protection Agency (EPA) discovered that diesel engines sold by VW in the US contained a defeat device, that is, a device "that bypasses, defeats, or renders inoperative a required element of the vehicle's emissions control system", as defined by the Clean Air Act. According to US officials, the defeat device software was able, by means of sensors, to detect when the car was being tested. That is, the software included instructions that activated equipment that reduced emissions by adjusting catalytic converters and valves when the car was being tested. The same software turned such equipment off during regular driving, possibly to save on fuel or to improve the car's performance. This meant that the emissions from the cars increased above legal limits, even 40 times the threshold values.

So, in this case we have an example of software that is causally efficacious; the defeat device controlled the VW engines so that they would (seem to) pass the EPA test. Hence, the device is an agent according to the definition of causal agency. The case is fitting here because it is analogous to Bostrom's disaster scenario insofar as it involves software that harms humans.

Although the defeat device fits causal agency, the agency of the defeat device has not been emphasized in discussion of the case and no one (to our knowledge) has suggested that the defeat device was responsible for the fraud though it was an essential element in making the fraud possible.

2.2 Intentional agency

The second type of agency involves the capacity for intentional actions, that is, intentional agents are entities that act intentionally. In traditional accounts, only humans can have intentions. Since intentions are seen as mental states, artifacts do not, strictly speaking, have intentional agency. Computer scientists and others sometimes attribute intentions to artifacts but such attributions are metaphorical; artifacts are spoken of as if they had intentions. Some suggest that at some pivotal moment in the future, AI artifacts might come to have something comparable to human intentions but this is highly speculative. Intentionality in artifacts is only metaphorical. Hence, if someone were to say that the VW defeat device was an agent in bringing about the fraud, this would either mean causal agency or would have to be interpreted as intentional agency metaphorically.

Causal agency and intentional agency share the element of causal efficacy, but in Intentional Agency, the agent's intentions begin the chain of causality. Importantly, intentions and intentional action are linked to responsibility though the connection is complex. In ethical and legal contexts, the presence of particular types of intentions determines the ascription of responsibility.

Returning to the VW case, in the blame game that ensued after the fraud became public, top management and the engineers were targeted as the entities that were possibly responsible for the fraud. As said before, causal agency is not sufficient to initiate a discourse on responsibility. For that, intentional agency is needed, which is why the focus in the VW case has been on humans, i.e., VW top management and engineers. Top management claims that the decision to use the defeat device was made by the engineers once they realized that the engines on which they were working would never meet the EPA standards without significant improvement (i.e., investments by the company). Allegedly,

not wanting to be bearers of bad news to their higher-ups, the engineers handled the problem on their own, keeping the engines as they were, but adding the defeat device (Smith and Parloff 2016). On this account, top management may not have had the intention to break the law, though the engineers did.

Nevertheless, in the public debate about the case, intentional actions of top management come clearly into play. Top management acknowledges that they specified both goals on which the engineers acted (to achieve a particular level of performance for the car and to meet the EPA standards). Intentionally setting these goals and intentionally creating a corporate culture in which engineers feared the consequences of failure (and did not want to tell top management that these goals could not be met) can be seen as setting off the sequence of events that led to the fraud. The point is that the issue of responsibility depends not just on causal sequences but on intentions and intentionality.

2.3 Triadic agency

Causal agency and intentional agency each capture something important about how states of affairs in the world are produced. However, neither alone adequately captures the full story. For example, in the VW case, causal agency covers only the causal efficacy of the defeat device but not the intentionality of top management and the engineers. Intentional agency covers the intentionality and causality of the humans, but not the contribution of the artifact that the humans used to realize their intentions. The defeat device played a role here both in shaping the intentions of the engineers and in making the fraud possible. Although a simple combination of the two might seem to solve the problem, it would not capture how the two work together, that is, how the intentionality of human actors and the efficacy of artifacts interact with one another to produce results like the VW fraud.

We propose a combination that we refer to as triadic agency. We are not the first to develop a multi-component account of agency: the above-mentioned ANT is an example of a account involving a network of components. Our proposal is to use triadic agency to analyze events involving technological artifacts in a way that draws attention in particular to users, designers, and artifacts—the most powerful agents in this kind of event (Fig. 1).

Triadic agency is especially helpful in sorting out responsibility.

When humans act with artifacts to achieve goals:

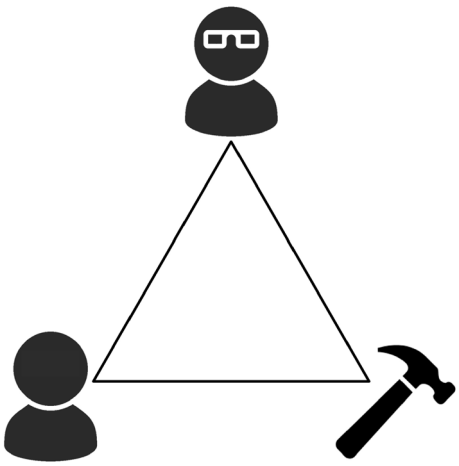

Fig. 1 The triadic agency. From bottom left anti-clockwise: the user, the artifact, the designer

- The *user* (or users) wants to achieve a goal and delegates the task of achieving that goal to the designer.[1]
- The *designer* (or designers) creates an artifact in order to achieve the goal.
- The *artifact* provides causal efficacy necessary to achieve the goal.

In the VW case, top management (representing the company) were the users, they had the goal of creating a car that would meet EPA standards while also meeting certain performance standards. The engineers were the designers—they were tasked with creating a car that would fulfill the goals of the top management and they did so by creating an artifact that would achieve the goal. They created the defeat device. All three together contributed to the achievement of the goal to pass the EPA test while also meeting performance standards and all three were essential to producing the emission fraud. All three are part of the agency of the illegal action.

This triadic account of agency allows us to identify agency in producing states of affairs while at the same time acknowledging that it is neither humans nor artifacts alone that do this. Humans and artifacts work together, with humans contributing both intentionality and causal efficacy and artifacts supplying additional causal efficacy. When users delegate to designers, they do so with the intention to achieve their goal and when designers accept the task,

[1] The term 'user' should be understood broadly here to refer to individuals or groups of individuals that commission the development and the deployment of an artifact for their purposes, i.e., to achieve their goals. In other words, the users in our Triadic Agency are "commissioning users", possibly distinct from the "end users" of the artifact, to whom the term 'user' usually refers to in other contexts, i.e., software engineering.

they intend to complete it, that is, they intentionally create artifacts that will achieve the delegated goal.

To illustrate the value of adopting a triadic agency account of agency, we can use it to think through issues of responsibility and in particular issues of responsibility with regard to increasingly autonomous artifacts. We will not limit ourselves to current technology but will show how triadic agency addresses issues of responsibility when it comes to fully autonomous futuristic artifacts.

3 Agency and responsibility in autonomous futuristic AI

In the current VW case, allegedly the engineers intentionally created the defeat device. The link between their intentions and the defeat device is direct: the defeat device is comprised of certain pieces of code, and that code was written by the engineers. Such directness may not hold in the future, since technologies are becoming more autonomous with humans delegating more and more complicated tasks to intelligent machines (e.g. software agents, robots). This raises difficult issues in the attribution of responsibility, because human beings will rely more and more on the intricate causal efficacy of these machines to accomplish their tasks. What if technology reaches a level where humans only need to specify a goal and machines are able to write the code to reach it? The value of the triadic account of agency can be demonstrated by addressing the challenges of responsibility ascription posed by the use of these emerging technologies and futuristic imaginations of them.

Imagine a futuristic scenario in which VW replaces its engineers with an advanced AI that is capable of designing software. Imagine further that this AI is given the goal of passing the EPA test without making the car more costly. Suppose the advanced AI notifies the VW management when the goal has been reached. The car passes the test with flying colors. Later the EPA discovers that the advanced AI has developed and embedded in the car's software instructions to control the engine in essentially the same way in which the 2015 defeat device operated. In test mode, catalytic converters and valves (or whatever equivalent parts there are in a futuristic VW engine) adjust so that the car passes the test, only to return to a non-EPA compliant mode once the test is over. If this deception were discovered, how would agency be understood and how would responsibility be ascribed?

3.1 The triadic agency in the futuristic case

Although some might argue the advanced AI is responsible ethically and/or legally, that would place responsibility where it would make no sense and do no good. This is shown by the triad analysis. Top management is still the user, specifying the goals it wants to achieve and the defeat device is still the artifact. However, the designer is now the advanced AI to which top management delegates its goals. The advanced AI, rather than human engineers, develops the defeat device. Even in this futuristic case, the user (VW management), the designer (the advanced AI), and the artifact (the defeat device) together constitute the agency that produces the fraud.

3.1.1 The artifact

The agency of the artifact requires the least analysis. The defeat device has causal agency: it is causally efficacious in the production of the fraud. It does not have intentional agency because it does not have any intentions. However, since the device plays a role in producing the fraud via its causal efficacy, we claim that it is part of the agency of the fraud (i.e., triadic agency).

3.1.2 The designer

The advanced AI produced the instructions that constitute the defeat device and its causal efficacy in the fraud. The execution of these instructions counts as cheating the test in the context of the EPA rules. Did the advanced AI produce those instructions intentionally? Certainly not in the same sense that the engineers in the current VW case intentionally produced the defeat device. The engineers in the current case are human beings with intentions (with Intentional Agency) while the advanced AI of the future is software running on a computer. Does this software have intentional agency? Some AI researchers might argue that if an AI is advanced enough its acts could be classified as intentional agency, but since it consists of software running on a computer, we argue that it is still in the realm of causal agency. To suppose that futuristic AI will have intentions in the same way as humans do today is to make a leap over the ontological chasm between computational artifacts and sentient beings. We will return to this chasm later. For now, we treat the advanced AI as having only causal agency.

Since advanced AI has only causal agency, it cannot have malicious intent or be negligent, both of which require intentionality. If either question were raised, the focus would quickly turn to the intentionality of those who had conceived, designed, deployed, or authorized the use of the advanced AI. Since the advanced AI has no intentions, responsibility cannot be ascribed to it. However, as with the defeat device, the advanced AI is part of the triadic agency of the fraud.

3.1.3 The user

As in the current case, in the futuristic case, top management is the user. It uses the defeat device designed by the advanced AI to reach the goal of passing the EPA test. Top management delegated this goal to the advanced AI. As with the artifact and the designer, the user is part of the agency that produces the emission fraud. However, top management is different in that, unlike the artifact and the designer, its members have intentions; they have intentional agency. Because of this, top management would be held responsible, and what they would be held responsible for would be dependent on the nature of their intentionality. That is, depending on the details of the case, they might be considered ethically reprehensible, negligent, strictly liable or to have committed some other complex infraction. As the only entity in the triad with intentional agency, was top management completely responsible for the fraud? Because of its reliance on the advanced AI, instead of jumping to such an ascription of responsibility, the relationship between top management and the advanced AI needs to be explored further. To start, we can ask: what was top management thinking when it delegated the achievement of its goal to the advanced AI? Answering this question brings into focus new relevant entities and a new triad: the designers of the advanced AI, and the triad for the creation of the advanced AI.

Suppose that top management hired a team of software engineers to create an advanced AI for car design and manufacture. In the triad for the creation of the advanced AI, top management is again the user, human software engineers are the designers, and any number of futuristic hardware and software tools would be the artifacts that facilitate the production of the advanced AI. Top management still plays the role of the user, this time with the goal of creating an advanced AI that could solve the problem of passing the EPA test. In this triad, triadic agency (the combination of user, designer, and artifact) produces an advanced AI capable of solving the problem of the EPA test. Here an analysis of the nature of responsibility for creating an advanced AI capable of producing something illegal would make sense.

In this triad for creating the advanced AI, the question of responsibility focuses on top management and the human designers. Since the VW involves a legal infraction, the question is whether and/or how each thought about the possibility of an illegal solution to the challenge of meeting the EPA test. Legal responsibility might depend on the answer to that question. For example, the user could have had malicious intentions or could have been negligent if it had specified the goal (to the human designers) to build an advanced AI that would solve problems regardless of the legality of the solutions. Or the designers might have disregarded the user's specification that the advanced AI be made only to do what was legal or the designers might have negligently failed to inquire whether the AI would be used in contexts in which legal requirements had to be met.

We will not solve the responsibility ascription in this futuristic case here. Our point is that even in the futuristic case, the ascription of responsibility would involve an inquiry into the behavior of the human components of the triad. We have separated agency and responsibility in the sense that the three components of the triad are all part of the triadic agency that produces the fraud though only the users are responsible because they are the only components that have intentional agency (the capacity for intentional action).

Why, it might be asked, is responsibility restricted to the humans in the triad? The answer is not that responsibility has some mysterious metaphysical meaning. Rather our answer is that responsibility is a mechanism for motivating people to behave in particular ways. This applies to both ethical and legal responsibility. When it comes to legal responsibility, lawyers and legislators often acknowledge that laws are intended to shape behaviour. Ethical responsibility also has a similar logic. Social notions of responsibility function to pressure individuals to act in ways that are good for society. In other words, when individuals believe they are responsible they are more likely to behave responsibly. In the case of technological mishaps, responsibility has to be placed on those entities who have the capacity to ensure or increase the likelihood that bad things will not happen.

We think agency and responsibility should be separated in the sense that agency is triadic while responsibility is always ascribed to humans. Of course, the kind of responsibility humans bear is complicated because of the delegation to artifacts and because of the complexities of negligence, malicious intentions, strict liability, etc.

3.2 Triads without humans: can non-humans have intentional agency?

Could there be a triad in which no human and no intentional agency is involved? How would the ascription of responsibility work in that case? In order to pose these questions we have to imagine a futuristic world well beyond the one we have been imagining. As an illustration, we can take Bostrom's futuristic depiction of a superintelligence that gives itself goals and designs solutions for those goals. Imagine a version of the VW fraud case that involves a superintelligence with human-like desires and intentions and able to take over the world. We call this an *endpoint* VW case, because AI futurologists see such superintelligence as an artifact at the endpoint of technological development. The main difference between this scenario and the earlier futuristic case is that the superintelligence occupies all three vertices of the triad: it would be the user with the goal of passing the EPA test, the designer with the capabilities to conceive

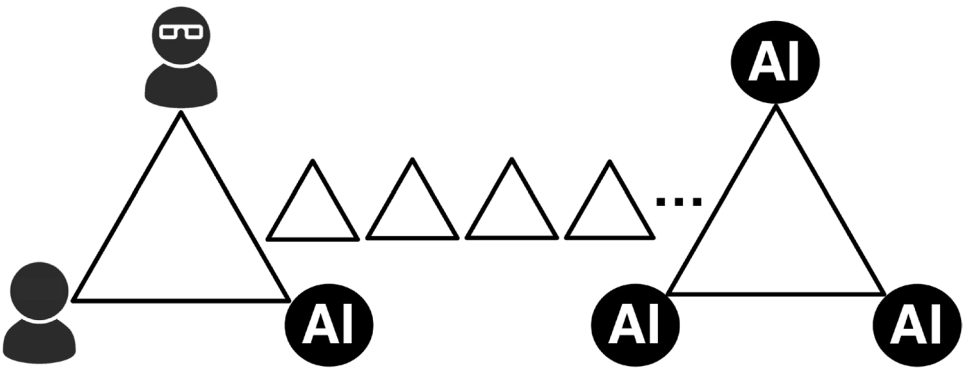

Fig. 2 The chain of triadic actions in technological development always involves human decisions at a certain point

and implement a solution meeting the EPA standards, and would embody that solution, that is, it would be the artifact. The triad has collapsed into one superintelligent entity that, presumably, is and does everything, and humans have become completely irrelevant.

Earlier we said we would return to the question whether futuristic AI might have something that would qualify as intentional agency. AI researchers treat AI artifacts as causally efficacious agents, that is, causal agents. However, in their speculations about the future, they imagine that AI artifacts will include computational elements that will make them agents in the way humans are (Bostrom 2014; Omohundro 2016; Yampolskiy 2016). Without worrying too much about how computation can create something equivalent to intentions, these AI researchers slip into thinking of futuristic AI as having intentional agency. Indeed, futuristic AI scenarios import the full mental complexity of humans to their AI artifacts talking as if AI can have drives, interests, goals, as well as intentions. Here they seem to get trapped in the metaphorical meaning of intentional agency.

AI is computational, whereas intentions are not, that is, the two are ontologically different. To claim otherwise presumes that computationalism is correct, that is, that something computational could be the same as a human mental state, i.e., superintelligence would have intentions. We do not claim that technological development could never take this path because we do not have decisive evidence against computationalism. On the other hand, neither can AI futurologists provide evidence in support of the possibility that AI artifacts can have intentions like humans do.

Whatever endpoint in the future we can imagine—be it one in which AI has intentional agency or not—ascriptions of responsibility will follow the sequence of human steps that will have led to that endpoint. As in the futuristic VW case, humans would have had to make decisions to design AI technology in particular ways and would have had to delegate operations to it. Those humans would be responsible for the AI behavior. Yes, the more developed the AI will become, the more distant human decision making will be from the execution of operations by the AI. Still, however distant the human decisions, they will have to have been part of the process at one point or another. Our triadic account helps keep track of the agencies and responsibilities in this process (Fig. 2).

4 Conclusion

The account of agency that we have presented is intended to frame agency in a way that acknowledges the combination of contributions of users, designers, and artifacts to produce states of affairs. The account recognizes that the agency at work in the production of states of affairs in the world is neither singular nor exclusively human. Although users, designers, and artifacts each constitute agency, none of them alone can achieve any particular technological state of affairs in the world.

The question of agency arises most often when it comes to responsibility, for example, in questions such as who is responsible for the VW emission fraud. Our account of agency provides a basis for analyzing responsibility while at the same time clarifying the distinction between agency and responsibility. The account purposely selects the role of users and designers in order to deal with issues of responsibility. In this way, we are able to represent the key contributors to states of affairs without slipping into an oversimplified view according to which, because of complexity, no human can be held responsible.

The account is applicable to both present day cases (as we have illustrated through the VW case) and futuristic possibilities. In the former, with human users and designers, we can easily detect the locus of responsibility in connection with their intentions, i.e., intentional agency. In futuristic cases, in which AI technology replaces humans in the role of the designer choosing how to achieve human goals, the distance between human intentions and artifact operations increases. Still, the triadic account allows us to trace human responsibility.

In more distant, imagined futuristic cases, no humans are presumed to be involved in a triad, that is, AI technology

occupies the roles of user and designer. Attribution of responsibility then calls for a search into the circumstances that led to the implementation of such a technology. Has it been designed by humans or by other machines? We need to trace back until we find a triad with human users or designers to be able to engage in a discourse on responsibility. The distance between the AI technology with its humanless triad and the point at which humans delegated goals to non-humans may be significant. This would make it difficult (but not impossible) to attribute responsibility to humans for an AI mishap, at least in the form of negligence (if not malicious intent).

In a scenario of superintelligent AI, negligence on behalf of humans would be based on the delegation of goals to AI technology without setting adequate boundaries on its behavior. However, we are not interested in the blame game that might ensue among humans enslaved by machines in an imagined future. We instead propose our triadic analysis of agency to sort out what happens *now* when humans use technological artifacts to bring about changes in the world, especially when, as in the VW fraud case, humans are behaving badly and in ways that are harmful to other humans.

A careful triadic analysis sheds light on the path of technological development currently underway and shows that, however complex the technology, responsibility still lies with humans. By pointing this out, our effort aims at keeping responsibility where it will do the most good: to encourage humans that design and deploy AI technology to anticipate the role of the technology in producing states of affairs.

References

Allen C, Wallach W (2012) Moral machines: contradiction in terms or abdication of human responsibility. In: Bekey GA, Lin P, Abney K (eds) Robot ethics: the ethical and social implications of robotics. The MIT Press, Cambridge, pp 55–68

Barrat J (2013) Our final invention: artificial intelligence and the end of the human era. Macmillan, London

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Castelfranchi C, Falcone R (2003) From automaticity to autonomy: the frontier of artificial agents. In: Hexmoor H, Castelfranchi C, Falcone R (eds) Agent autonomy. Springer, New York, pp 103–136

Colburn TR, Shute GM (2008) Metaphor in computer science. J Appl Logic 6(4):526–533

Dignum F (1999) Autonomous agents with norms. Artif Intell Law 7(1):69–79

Dormoy JL, Kornman S (1999) Meta-knowledge, autonomy, and (artificial) evolution: some lessons learnt so far. In: Varela FJ, Bourgine P (eds) Toward a practice of autonomous systems. The MIT Press, Cambridge, pp 392–398

Dowd M (2017) Elon Musk's billion-dollar crusade to stop the A.I. apocalypse. Vanity fair, April 2017. http://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x. Accessed 27 April 2017

Ferber J (1999) Multi-agent system: an introduction to distributed artificial intelligence. Addison-Wesley Longman, Boston

Floridi L (2008) Artificial intelligence's new frontier: artificial companions and the fourth revolution. Metaphilosophy 39(4–5):651–655

Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14(3):349–379

Franklin S, Graesser A (1997) Is It an agent, or just a program? A taxonomy for autonomous agents. In: Müller JP, Wooldridge MJ, Jennings NR (eds) Intelligent agents iii agent theories, architectures, and languages. ATAL 1996. Lecture notes in computer science (1193). Springer, Berlin

Hotten R (2015) Volkswagen: the scandal explained. BBC news, December 10, 2015. http://www.bbc.com/news/business-34324772. Accessed 27 April 2017

Jennings NR, Sycara K, Wooldridge M (1998) A roadmap of agent research and development. J Auton Agents Multi Agent Syst 1(1):7–38

Jennings NR, Norman TJ, Faratin P, O'Brien P, Odgers B (2000) Autonomous agents for business process management. Appl Artif Intel 14(2):145–189

Johnson DG, Noorman M (2014) Artefactual agency and artefactual moral agency. In: Verbeek P, Kroes P (eds) The moral status of technical artefacts. Springer, The Netherlands, pp 143–158

Johnson DG, Verdicchio M (2017) Reframing AI discourse. Minds Mach. https://doi.org/10.1007/s11023-017-9417-6

Latour B (1996) On actor-network theory: a few clarifications. Soziale Welt 47(4):369–381

Law J, Hassard J (1999) Actor network theory and after. Wiley, NJ

Lin Z, Broucke M, Francis B (2004) Local control strategies for groups of mobile autonomous agents. IEEE Trans Autom Control 49(4):622–629

Maes P (1990) Designing autonomous agents. In: Maes P (ed) Designing autonomous agents: theory and practice from biology to engineering and back. The MIT Press, Cambridge, pp 1–3

Malle BF, Scheutz M (2015) When will people regard robots as morally competent social partners? In: 24th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 486–491

Noorman M, Johnson DG (2014) Negotiating autonomy and responsibility in military robots. Ethics Inf Technol 16(1):51–62

Omohundro S (2016) Autonomous technology and the greater human good. In: Müller V (ed) Risks of artificial intelligence. CRC Press, Boca Raton, pp 9–27

Pfeifer R (1995) Cognition—perspectives from autonomous agents. In: Steels L (ed) The biology and technology of intelligent autonomous agents. Springer, Berlin, pp 128–164

Powers TM (2013) On the moral agency of computers. Topoi 32(2):227–236

Sayes E (2014) Actor network theory and methodology: just what does it mean to say that nonhumans have agency? Soc Stud Sci 44(1):134–149

Schlosser M (2015) Agency. In: EN Zalta (ed) The stanford encyclopedia of philosophy, fall 2015 edition. https://plato.stanford.edu/archives/fall2015/entries/agency/. Accessed 27 April 2017

Smith G, Parloff R (2016) Hoaxwagen. Fortune, March 15, 2016. http://fortune.com/inside-volkswagen-emissions-scandal/. Accessed 6 November 2017

Steels L (1995) When are robots intelligent autonomous agents? Robot Auton Syst 15(1–2):3–9

Stone P, Kaminka GA, Kraus S, Rosenschein JS (2010) Ad hoc autonomous agent teams: collaboration without pre-coordination. In: proceedings of the twenty-fourth AAAI conference on artificial intelligence, pp 1504–1509

Tentori M, Favela J, Rodriguez MD (2006) Privacy-aware autonomous agents for pervasive healthcare. IEEE Intell Syst 21(6):55–62

Weiss G (1999) Multiagent systems: a modern approach to distributed artificial intelligence. The MIT Press, Cambridge

Wooldridge M (2002) An introduction to multiagent systems. Wiley, NJ

Yampolskiy RV (2016) Utility function security in artificially intelligent agents. In: Müller V (ed) Risks of artificial intelligence. CRC Press, Boca Raton, pp 115–140