Strictly Human: Limitations of Autonomous Systems

Soltanzadeh, S.

DOI
10.1007/s11023-021-09582-7

Publication date
2022

Document Version
Final published version

Published in
Minds and Machines

Minds and Machines (2022) 32:269–288
https://doi.org/10.1007/s11023-021-09582-7

GENERAL ARTICLE

Strictly Human: Limitations of Autonomous Systems

Sadjad Soltanzadeh[1]

Received: 14 July 2021 / Accepted: 6 November 2021 / Published online: 13 November 2021

Abstract
Can autonomous systems replace humans in the performance of their activities? How does the answer to this question inform the design of autonomous systems? The study of technical systems and their features should be preceded by the study of the activities in which they play roles. Each activity can be described by its overall goals, governing norms and the intermediate steps which are taken to achieve the goals and to follow the norms. This paper uses the activity realist approach to conceptualize autonomous systems in the context of human activities. By doing so, it first argues for epistemic and logical conditions that illustrate the limitations of autonomous systems in tasks which they can and cannot perform, and then, it discusses the ramifications of the limitations of system autonomy on the design of autonomous systems.

Keywords System autonomy · Design · Activity realism · Objectivity · Regulation

1 Introduction

Autonomous systems can be defined as systems which are capable of acting upon the world independently of real-time human control, such as autonomous vehicles or autonomous military robots; both still in the development phase. The suitability of the term 'autonomous' to refer to technical systems can of course be questioned. This is because current machines are not autonomous in the sense in which the term is used in moral philosophy; i.e. they do not have the capacity to reflect on and understand reasons for actions and cannot be held morally or legally responsible for their impacts on their surroundings (Nyholm, 2018; Purves et al., 2015). However, here a basic sense of autonomy is meant which Sartor and Omicini (2016) would classify as a non-cognitive conception of autonomy. In this sense, 'a system is autonomous to the extent that it can accomplish a task by itself, without external directions. Once the system starts its activity, there is no need for other human or

✉ Sadjad Soltanzadeh
s.soltanzadeh@asser.nl

[1] T.M.C. Asser Institute, University of Amsterdam, The Hague, Netherlands

artificial agents to monitor its behaviour and govern its functioning' (Sartor & Omicini, 2016, p. 39). So, the sense in which the phrase 'autonomous systems' is used here does not carry any metaphysical assumptions about the internal properties of the systems in question. The phrase' autonomous systems' is used as an umbrella phrase to refer to a range of technical systems. In fact, the scope of the systems covered here is wider than systems that use Artificial Intelligence (AI). Arguments of this paper apply to less complex systems, such as the landmine or the automatic door as well.

One of the central questions which has guided the research into autonomous systems is whether these systems can replace humans in their activities. The replaceability of humans by autonomous systems can have diverse economic, cultural, societal, cognitive and personal identity ramifications (Aizawa, 2013; Coeckelbergh, 2021; Gasser, 2021; Haselager, 2013). Currently, most works which address the replaceability of humans by autonomous systems do so by focusing on the normative question of whether the performance of certain human activities should be delegated to autonomous systems; whether it is *desirable* to delegate the performance of activities to autonomous systems; whether there are any decisions or actions which, for moral or legal reasons, should be made or performed only by humans. For instance, some have argued that autonomous systems are likely to lead to responsibility or retribution gaps, and as such, it is morally wrong to introduce autonomous vehicles and autonomous weapon systems into society which will inevitably need to make morally significant choices (Danaher, 2016; de Jong, 2019; Matthias, 2004; Roff, 2014; Sparrow, 2007, 2016). In the case of the military application of AI systems, some have argued that machines are not able to discriminate between civilians and combatants, and as such, the development and use of autonomous weapon systems should be restricted or banned altogether (Krupiy, 2015; Sharkey, 2010). On the other hand, some others have argued that autonomous systems are likely to cause less harm than humans, and that autonomous systems should not be held to higher standards than their human counterparts (Schulzke, 2011; Strawser, 2010).

Here, however, the replaceability of humans by autonomous systems is addressed from the epistemic and logical perspectives. The question is not whether autonomous systems *should* but *are able to* replace humans in their activities. In this respect, this paper provides a similar insight as works which have discussed the general philosophical limitations of autonomous systems and AI. Most notably, Dreyfus (1993) has refuted the epistemological assumption in some AI programs that human behaviour can be formalized and that this formalism can be used to artificially reproduce human behaviour. Dreyfus rejects this assumption by arguing that human behaviour cannot be expressed by 'laws of behaviour' which can later be embodied in a computer program. Similarly, Suchman (1987) has made a methodological argument that human behaviour and the interactions that humans have with their environment cannot be formalized by using symbolic models and methods. Searle (1980), on the other hand, has provided a metaphysical argument to show that even if one day AI products can exhibit behaviours which are indistinguishable from human behaviours, these machines still do not possess consciousness and intentional states, such as beliefs and desires. These epistemic and metaphysical arguments show non-normative limitations of autonomous systems.

This paper uses a recent approach developed in the metaphysics of technology, namely, activity realism (Soltanzadeh, 2019), in order to illustrate specific epistemic and logical conditions which need to be fulfilled in order for autonomous systems to be able to replace humans in their activities. The epistemic condition concerns whether the performance of certain human activities can possibly be taught to autonomous systems. The logical condition is whether delegating the performance of these activities to autonomous systems is consistent with the activities' logical structure; whether human users can remain engaged in certain activities if some tasks in those activities are performed by autonomous systems. Based on these two conditions, activities can be categorised into four groups: those which satisfy neither, one or both of the conditions. The majority of this paper's analyses are on two extreme types of activities which satisfy neither or both conditions. Using examples taken mostly from activities in which vehicles are used, the next four sections offer a philosophically rich analysis of activities in which autonomous systems can and cannot replace humans, and the final section discusses some implications of this analysis for the design and regulation of autonomous systems.

2 Activity Realism and Autonomous Systems

To study limitations of system autonomy through logical and epistemic lenses, a philosophical approach is required that is fit for this task. The chosen approach needs to connect theory with practice by generating concepts and distinctions that are fruitful for studying practical issues regarding the place of autonomous systems in our practices. Not all concepts are useful, and practical issues may be insensitive to some abstract distinctions. For example, the distinction made between planets and dwarf planets can be conceptually interesting, but the scientific theories that are used to study celestial bodies are insensitive to this distinction. For instance, when the classification of Pluto was changed from a planet to a dwarf planet, astrophysicists did not need to use different theories to calculate the trajectory of the New Horizons interstellar probe. Similarly, it has been argued that the distinction between artefacts and non-artefacts is irrelevant for achieving practical goals, and artefacts are not necessarily more philosophically interesting than non-artefacts (Soltanzadeh 2019).

Practically useful concepts and categorisations are those that are developed in relation to activities of reflective beings, such as humans (Soltanzadeh, 2019). Categorisations and distinctions that are developed in relation to context-independent properties of objects are not useful for practical, functional and normative studies. The approach which categorises objects in relation to human activities is called 'activity realism' (ibid.). Hence, to study the logical and epistemic limitations of autonomous systems, the activity realist approach has been chosen here. Other approaches in the metaphysics of technology are theoretically insightful, but they may not be fruitful for practical decision-making and evaluative purposes, as they tend to categorise objects and systems based on their context-independent properties (ibid.).

In activity realism, we start our identification of reality from activities of reflective beings, such as humans. What is taken to be fundamentally real are activities. Objects are identified in relation to their place in human activities; they are not identified by their intrinsic properties (Soltanzadeh, 2019). A paper bag, for instance, acquires different identities when it is used as a rubbish bin, a sick bag or wrapping paper. In activity realism, we cannot merely point at an object and claim to have identified an object in addition to the atoms and molecules which constitute it. Any practically useful identification of the object assumes a particular activity in which it can be used (ibid.). Hence, in activity realism the study of objects and systems is preceded by the study of activities.

Given that in activity realism objects derive their reality from the roles which they play in activities, the evaluation of objects must be sensitive to their role (Soltanzadeh, 2019). Since practical evaluations of an object are activity-dependent, the adjustments which are required to optimize an object to be used in a particular activity may make it better or worse for the performance of a different activity (ibid.). To make a paper bag a better bin bag, for example, it may need to become water resistant and tough. However, the first adjustment is unnecessary for a paper bag to be used as wrapping paper, and the second would only make the paper bag a worse wrapping paper, as fine wrapping papers may be more appealing.

One of the adjustments made to technical systems is adding autonomous features to them. Autonomous features are often added to pre-existing systems in order to improve their performance. Autonomous cars, for example, are meant to be an improved version of normal cars, which are already widely used in society.

As systems become more autonomous, more tasks are delegated from humans to machines. However, as the performance of certain actions are taken away from humans and delegated to autonomous systems, the nature of user engagement changes. This change may in some cases be desirable as users will be able to focus on other tasks. In fact, one of the marketing features of autonomous vehicles will be that they allow users to engage in other activities, such as reading books or checking their mobile phones, while the vehicle performs the task of driving.

But removing the possibility of user engagement may also limit the activities in which the technical system can be used, as autonomous systems may not be able to replace humans in some activities. As such, it is imperative to study the conditions which need to be met in order for autonomous systems to be able to replace humans. This study can highlight the types of activities which autonomous systems can and cannot perform. To conduct this study, different types of involvement in activities and the logical structure of activities need to be first examined in more detail.

3 General Norms, Activity-Dependent Goals and Momentary Goals

Linguistically, *engagement in activities* and *performing activities* are sometimes used interchangeably. But for the sake of the arguments of this paper, a distinction can be made between engagement in an activity and performing actions which are required to achieve the goal of an activity. One can be engaged in the activity of using a vehicle to go from a particular place to a destination by performing the act

of driving oneself or by delegating the act of driving to another driver, be it a human taxi driver or an intelligent technical system. One who is engaged in the activity of commuting from home to work is the passenger, not the taxi driver or the vehicle. One who performs the act of driving the vehicle is the one who needs to have the required skill and knowledge to be able to drive a vehicle from the passenger's home to their work. Although autonomous systems can *perform* a series of actions, the only entity who is *engaged* in the activity is the user.

In addition to this distinction, further clarifications about the structure of activities need to be made before the main arguments of this paper can be laid out. In general, activities can be studied through the general norms governing them, their goals and the actions required to bring about the goals by observing the norms. Each of these can be separately explicated.

3.1 General Norms

To perform an activity, each person may consider a series of general norms which are in place to govern the activity. Road rules are an example of general norms. Regardless of why a person is using their vehicle on a public road, their driving style, the routes which they follow or the model of their vehicle, they should stop at red lights or drive on the right side of the road.

General norms are often, but not always, prescribed by authorities. Health and safety regulations, legal rules and moral codes are examples of general norms set by authorities. These general norms are often context-sensitive, and they remain valid throughout the context. Whether one is required to drive on the left or right side of the road, for example, depends on the country in which they are driving.

However, not all general norms are decided by authorities. Some general norms may be conventional or personal. Consider a car user who has a strong sense of environmental responsibility and wishes to minimize their greenhouse gas emissions on every trip. Minimizing greenhouse gas emissions is a personal value, as other users may not respect it. But it may be upheld over a wide range of activities in which the user engages and work as a general norm for that particular user. As another example, consider a user who wishes to exhibit altruistic character traits while driving their car. They may do so by slowing down their car to allow other cars to join the traffic or change lanes. Showing altruistic character traits while driving is also a personal norm, but it may be upheld over a wide range of activities by a user and work as a general norm for them.

Therefore, what makes something a general norm for a specific user is the fact that the user considers that norm in decisions which they make over a wide range of activities. General norms, therefore, are frameworks through which goals are achieved. They work in the background and govern activities.

3.2 Activity-Dependent Goals

Activity-dependent goals are the reasons which motivate engagement in activities. In the context of using a vehicle, for instance, a user may want to use their vehicle purely as a means in order to go from home to work. Or they may want to look for their escaped dog. Or they may want to perform a burnout, or rather aimlessly go for a leisurely drive. Each of these goals is an activity-dependent goal as they motivate engagement in an activity. These goals vary from one individual to another and can dynamically change for one individual.

Successful completion of an activity requires the achievement of the goal of the activity. The activity of going from home to work, for instance, is successfully completed when the user arrives at work. The activity of leisurely driving is successfully completed when the user has had a joyful drive.

Activity-dependent goals are often set by users. Users have the freedom to decide which activities they wish to engage in. However, authorities may also regulate and restrict certain activities. For example, users are not allowed to use their vehicle to purposefully inflict harm on others.

3.3 Momentary Goals

A user cannot follow their general norms or achieve activity-dependent goals simply by sitting and waiting for the norms and goals to realize. More specific intermediate steps need to be set in order for general norms and activity-dependent goals to be realized. These intermediate steps, which can be called *momentary goals*, shape the momentary actions which are taken by the user. For example, when a user aims to go from home to work, they will need to take many momentary actions which jointly contribute to the achievement of the general norms and activity-dependent goals. At each intersection, they need to decide to turn left, right or go straight in order to arrive at work rather than at other locations. They need to decide which route to take and dynamically adjust the route in case of traffic congestion or road closures. They also need to constantly decide if and when to overtake other vehicles on the road, when to accelerate, push the brakes, change the gears or give way to other vehicles to join the traffic on a busy road. Finally, they need to consider general norms, such as road rules, and make momentary decisions which are consistent with these norms. Momentary goals, therefore, are dynamic and short-lived.

Momentary goals also have action-guiding roles. By setting momentary goals, general norms and activity-dependent goals can be broken into smaller, achievable steps, so that the achievement of these smaller steps leads to the achievement of general norms and activity-dependent goals. Momentary goals, in other words, can reveal *how* to perform an activity.

Momentary goals are meant to contribute to, and need to be justified by, general norms and activity-dependent goals. Different activities require different momentary goals, and while a momentary goal may be justified in one activity, it may not be justified in another. For instance, if the activity in which the user wants to engage is to go for a leisurely drive, they can justifiably slow down to look at the houses which are

recently constructed, or make a sudden stop to purchase an ice cream. However, these decisions would not be justified if the user wants to go from home to work as quickly as possible.

3.3.1 The Instrumental vs. Constitutive Roles of Momentary Goals

Performing an activity requires taking intermediate steps (achieving momentary goals) which contribute to the completion of the activity. However, the way momentary goals support activity-dependent goals can vary. Depending on the activity, momentary goals may be *instrumental* or *constitutive* of the activity. This point can be exemplified with two activities.

The first activity to consider is using a vehicle to go from home to work. The goal of this activity is to bring about a particular state of affairs; the state of affairs in which the user has arrived at work. This state of affairs, which represents the successful completion of the activity, is defined prior to the engagement in the activity. Before using the vehicle, the user knows what the world would like upon the completion of the activity. The success or failure of the activity can also be assessed solely by the changes made to the outside world; i.e. the whereabouts of the user at the end of the activity.

What defines engagement in the activity of using a vehicle to drive from home to work is the initial and the final states of affairs. The activity is defined by the changes which are made to the outside world. One is engaged in the activity of going from home to work if one is going from home to work. It does not matter if the person goes fast or slow. It does not matter if the person shows respect to other drivers or not. It does not matter which route is taken. In a nutshell, as long as the person gets to the work, it does not matter which or how momentary goals are set and followed in order to bring about the final state of affairs. This is because the activity is not defined by momentary goals, but by the initial and final states of affairs. Therefore, in the activity of driving from home to work, momentary goals are *instrumental* for the realization of the activity-dependent goal. From here on, activities such as using a vehicle to go from home to work are referred to as Type 1 activities.

Now consider the activity of leisurely driving. In the activity of leisurely driving, momentary goals play a *constitutive* role. This activity is performed to bring about a state of relaxation; its goal is about the quality of the experience of driving; it is not about bringing about a particular state of affairs. What defines the activity of leisurely driving is shown through the way momentary goals are set and followed. One is engaged in the activity of leisurely driving when one makes relaxed and often unpredictable decisions which are triggered by spur-of-the-moment feelings and curiosities. These relaxed and unpredictable momentary decisions *constitute* the activity of leisurely driving.

The constitutive role of momentary goals in activities such as leisurely driving means that if momentary goals are set differently, the nature of the activity can change. When relaxed decisions which are triggered by spur-of-the-moment feelings and curiosities are taken away from an activity and are replaced by constant accelerations, brakes and rapid turns of the steering wheel, the activity of

leisurely driving may turn into a different activity, such as the activity of torturing the passengers. In activities such as leisurely driving and torturing the passengers, therefore, whether one is successfully engaged in the activity depends on the way momentary goals are formed and pursued. From here on, this second group of activities are referred to as Type 2 activities.

4 Replaceable Activities

In the previous section, two examples demonstrated two different relations that momentary goals can have to activity-dependent goals. In this section and the next, the distinction between Type 1 and Type 2 activities is used to argue that Type 1 activities can and Type 2 activities cannot be replaced and performed by autonomous systems. First, further clarifications need to be made about what is required to perform an activity.

What does it mean to perform an activity? And what does it require for an entity, be it human, animal or an autonomous system, to be able to perform an activity?[1] Here the performance of an activity is defined as taking a number of intermediate steps which can jointly bring about the goal of the activity. The activity of changing a broken lightbulb, for example, requires separating the old bulb from the socket, grabbing the new bulb and screwing the new bulb in. Each of these steps needs to be broken into yet smaller steps. For example, to open the broken bulb, one may need to first bring a chair, stand on top of the chair, raise one of their arms, grab the bulb and twist the bulb counter-clockwise until the bulb is separated from the socket. Each of these smaller steps works as a momentary goal which is set in order for the activity-dependent goal of changing the lightbulb to realize. These momentary goals jointly make up the *plan of action* for changing a broken lightbulb.

Here it is worth clarifying that the activity of changing a lightbulb is a Type 1 activity, similar to the activity of using a vehicle to go from home to work. The goal of the activity of changing a lightbulb is to make changes to the outside world (i.e. to replace the old bulb with the new bulb). Also, in this activity the intermediate steps which are taken to perform the activity are instrumental for the realization of the goal of the activity. It does not matter what or how momentary goals are set, as long as they contribute to the completion of the activity. For example, it does not matter which stool is used to reach the bulb, who changes the bulb or how quickly the old bulb is unscrewed from the socket.

Overall, at least three conditions need to be met for an entity, such as an autonomous system, to be able to perform an activity: the physical, epistemic and logical conditions.

[1] Here the word 'entity' is used in its broad philosophical sense, as a general term for an arbitrary token of any kind, including humans and machines.

4.1 The Physical Condition

The first condition is for the entity to have the right physical capacities. In the case of autonomous systems, this condition is met when systems are designed with physical features which can be used to perform the task. To be able to change a lightbulb which is installed on a ceiling, a robot, for example, needs to be able to physically reach the lightbulb. One which is only 20 cm tall and is unable to expand or to fly, or to climb the walls or other furniture, cannot perform the function of changing a lightbulb, because it does not have the physical features to do so. Autonomous systems' physical capacities are restricted by available materials and production techniques. However, technical systems can sometimes far exceed humans in this respect. Technical systems can be designed to reach places where humans cannot reach or function in environments which are too dangerous for humans. So autonomous systems' physical capacities will not always hinder them from replacing humans in performing different activities.

4.2 The Epistemic Condition of Performance

The second condition which needs to be met for an entity to be able to perform an activity is an epistemic condition. To be able to perform an activity, the entity needs to have a plan of action for the performance of the activity. Having a plan of action is important because, as discussed earlier, the goals of an activity do not magically realize out of the blue. A number of intermediate steps, or momentary goals, are required to be taken for the activity-dependent goals to come true. Having a plan of action means being endowed with, or being able to devise, momentary goals whose realizations would lead to the realization of the activity-dependent goals. In the example of changing a lightbulb, it means knowing that changing a lightbulb requires smaller steps, such as unscrewing the broken bulb and screwing in the new one.

For autonomous systems to be able to replace humans, they need to have plans of action. So, for what types of activities can autonomous systems have plans of action? The answer to this question depends on the status of the intermediate steps required to perform the activity. It is much easier for autonomous systems to have a plan of action for an activity when the intermediate steps are objectively identifiable. Objectively identifiable here means something which can be referred to or perceived by sensory organs. This sense of objectivity is what Douglas (2004) identifies as the first mode of objectivity.

Objectively identifiable intermediate steps can be taught to autonomous systems. This is because autonomous systems can be equipped with sensors which can scan the environment, get feedback about their performance and make the necessary adjustments in order to successfully perform the intermediate steps. In fact, artificial sensors can surpass human sensory organs in the range of data that they can gather from the environment. Therefore, if the performance of an activity can be broken down into objectively identifiable steps, then it is not impossible for autonomous systems to learn to perform that activity.

Now the question is: what types of activities can be performed by following objectively identifiable intermediate steps? The intermediate steps required to perform Type 1 activities are objectively identifiable. This is because Type 1 activities are aimed at making changes to the objective world, and making changes to the objective world requires reliance on causal forces. In Type 1 activities, intermediate steps causally link the initial state of affairs to the final state of affairs. In other words, in these activities, intermediate steps, if followed correctly, would *cause* the achievement of the activity-dependent goals. That is why the momentary steps required to achieve the goal of a Type 1 activity can be derived from activity-dependent goals. For example, using a step or flying can help to reach a lightbulb which is installed on the ceiling. Grabbing a lightbulb and twisting it counter clockwise results in the separation of the bulb from the socket. So in Type 1 activities, the intermediate steps are objectively identifiable as they cause the states of affairs which represent the completion of these activities.

The fact that the intermediate steps required to perform Type 1 activities are objectively identifiable makes it possible for autonomous systems to have plans of action for these activities. In the activity of using a vehicle to go from home to work, for example, the intermediate steps which need to be set in order to bring about the activity-dependent goal can be objectively defined as following one of the routes which take the user from home to work. The autonomous vehicle which is used to perform this activity can get dynamic feedback about its position in relation to the destination, use the accelerator, brakes, gears and the steering wheel to navigate itself, and follow the routes which would lead to the destination. All of these steps are objectively identifiable, and if they are followed, the activity will be successfully completed.

4.3 The Logical Condition of Engagement

The third condition which needs to be met for an entity to be able to perform an activity is the logical requirement for engagement in the activity. As mentioned earlier, autonomous systems are meant to contribute to *human* activities. Even though they perform certain tasks for humans, the only entity who is engaged in an activity is the human user. The logical requirement stipulates that the definition of the activity should allow the transfer of the performance of the activity to others. To be able to perform an activity for a user, an autonomous system needs to preserve the activity in the logical sense so that the user remains engaged in that particular activity. Preserving the activity here means not spoiling the activity by ceasing it or turning it into a different activity. An example can clarify this point.

The religious activity of worshiping a deity is typically a human activity. The performance of this activity, say, in terms of hand gestures or words which need to be uttered, may be objectively identifiable. As such, autonomous systems can learn how to perform the required hand gestures and utterances. However, some religions do not allow the activity of worshiping to be delegated to others. In these religions, one needs to worship the deity oneself. Therefore, from a religious point of view, as soon as an autonomous system replaces a human in the performance of the activity

of worshiping, the human will not be engaged in the activity of worshiping the deity anymore. In other words, although autonomous systems may be implemented in religious contexts, they cannot replace the performance of the activity of worshiping for humans, because as soon as they do so, the activity of worshiping ceases and is replaced, say, by an entertaining activity or show business. The active participation of the human subject is a logical requirement for engagement in the activity of worshiping. This is how the activity is defined.

However, unlike the activity of worshiping, Type 1 activities are solely defined by the changes which need to be made to the world. In Type 1 activities, the momentary goals are only instrumental: it does not matter *who* performs the momentary goals or *what* momentary goals are achieved. As long as the final states of affairs are realized, the activity is successfully completed. As such, in Type 1 activities, delegating the performance of the intermediate steps to others does not spoil the activity. The activity of changing a lightbulb or that of driving from point A to point B, for example, can be delegated to others, such as friends, family members, taxi drivers, robots or autonomous vehicles. One can still achieve the goal of getting a lightbulb changed or going from home to work without performing the acts of changing the lightbulb or driving a vehicle oneself. Delegating these activities to other entities does not spoil them.[2]

To conclude this section, autonomous systems can fulfil the three conditions which are required for them to replace humans in the performance of Type 1 activities. The first condition is the physical condition. This condition can be met by designing systems with physical capacities appropriate for the task. The second condition is the epistemic condition. Autonomous systems can be equipped with the plan of action to perform Type 1 activities because performing these activities can be expressed in terms of objectively identifiable intermediate steps. The third condition is the logical condition. Autonomous systems can replace humans in the performance of the intermediate steps required for Type 1 activities without spoiling the logic of these activities.

5 Irreplaceable Activities

The above conclusions cannot be generalized to Type 2 activities, such as leisurely driving, where the quality of the experience of engaging in an activity is also important. In the case of Type 2 activities, there are epistemic challenges in replacing the human performance of the activity by an autonomous system. Replacing human

[2] Borgmann (1984) would argue that delegating activities to autonomous systems indeed spoils human activities. However, what is spoiled, in Borgmann's framework, is not the logical structure of the activity. Rather, he believes that there are intrinsic values in the performance of activities, and by delegating the performance of activities to technical systems, these intrinsic values are lost. These values, which in the activity realist framework are categorized as general norms, are lost by what Borgmann calls the 'device paradigm', which has resulted in the commodification of objects. As such, Borgmann's arguments are not limited to autonomous systems per se. For him, some intrinsic values are lost even when one uses a central heater to warm up their building, instead of chopping woods and feeding their chimney by hand.

performance of the activity may also violate the logical requirement for engagement in these activities.

5.1 The Epistemic Condition of Performance

One of the epistemic challenges in using autonomous systems to perform Type 2 activities is that it is hard, if not impossible, to supply autonomous systems with plans of action which can be used to bring about activity-dependent goals. This is a challenge because there is no objective recipe for performing Type 2 activities.

Consider the activity of leisurely driving as an example. There cannot be an objective plan for performing this activity. This activity cannot be broken down into smaller steps which would necessarily bring about the achievement of the activity-dependent goal of leisurely driving, which is enjoying the ride. Firstly, there are always individual differences in what makes a drive a leisurely one. While some may enjoy driving fast on highways, some others may prefer to drive around national parks, and others around the city centre. Secondly, even the person who wants to go for a leisurely drive may not know in advance what momentary decisions they will make to perform this activity. This is because even on one trip, a person's preferences can dynamically change. The person may begin by driving towards a quiet leafy suburb, but before reaching there they may decide to follow their curiosity and drive towards some newly built apartments, only to notice an ice cream store on the way to the apartments, stop, have an ice cream and directly return home without going to the quiet leafy suburb or visiting the new apartments. Despite these dynamic changes and not going to places where they initially wanted to go, the person can still enjoy a leisurely drive. In fact, following the initial plan might have resulted in a boring ride.

The above-mentioned epistemic challenge is due to the fact that in Type 2 activities, the realization of the goal of the activity cannot be assessed by the changes made to the outside world. Whether one has enjoyed what is meant to be a leisurely drive cannot be assessed by the changes made to the position of the vehicle during the activity of leisurely driving. Rather, the realization of the goals of Type 2 activities can only be subjectively assessed by the individual who is engaging in the activity.

Since the goals of Type 2 activities are not to make specific changes to the outside world, one also cannot rely on objectively identifiable intermediate steps to fulfil these goals. No particular worldly intervention can necessarily lead to the fulfilment of the goals of Type 2 activities. Consider leisurely driving again. Leisurely driving involves momentary, unpredictable decisions made in reaction to surrounding distractions. A mathematically infinite number of distractions can trigger a mathematically infinite number of unpredictable reactions. The challenge here is that any of these distractions can be justifiably pursued. So there are no defined sets of objective intermediate steps which can cause the realization of the goal of the activity. In type 2 activities, there is a non-causal and subjective connection between momentary goals and activity-dependent goals.

The subjective aspect of the performance of Type 2 activities has other ramifications as well. One of these ramifications is that from a third person perspective, it is impossible to distinguish between an activity which is performed to bring about a predefined state of affairs (a Type 1 activity) and an activity in which forming and performing momentary objectives are constitutive of the activity (a Type 2 activity). Imagine a person who after finishing work, enters their car, takes a particular route and stops at a cafe to order a drink. What was the activity in which the person was engaged? Although from a third person perspective, everything about the movement of the vehicle may be known, this knowledge is insufficient to determine the type of activity in which the person was engaged. The person might have been engaged in the activity of going to a predetermined cafe (Type 1), say, to meet with someone. Or they might have been engaged in the activity of leisurely driving (Type 2). It is possible that the unpredictable decisions which the person makes during the activity of leisurely driving guide the vehicle through the exact same route as purposeful decisions which need to be made in order to take the person to a particular location. Objective knowledge is not sufficient to determine the nature of an activity or distinguish between activities of different type. A drive which may objectively seem to represent an activity in which the user goes from a particular point to a destination may be a leisurely drive, a curious drive, a funny drive or a drive to frighten or torture the passengers. And in general, objective changes which are made to the world as a result of the performance of an activity may be the goal of that activity, as a Type 1 activity. But the same changes may be made for entertainment, religious or artistic reasons. In these latter cases, the exact changes may be irrelevant for the purpose of the activity.

The inherent subjectivity of Type 2 activities, therefore, poses an epistemic challenge for autonomous systems to replace the performance of Type 2 activities on behalf of their users. No other entity, even another human, may be able to perform these activities for others.

Here it should be acknowledged that some people may indeed take pleasure from delegating the task of leisurely driving to others. In fact, in some countries, one can ask a taxi driver to simply ‘take them for a drive’. The taxi driver may not even know the person, but the passenger may still enjoy the drive. Similarly, some people may willingly let autonomous vehicles take them for leisurely drives, in the same way that some people happily let algorithms stream random music for them on online platforms. However, this does not affect the epistemic argument made in this section. The argument does not rely on the premise that *no one* would be willing to allow autonomous systems or other humans to perform Type 2 activities for them. Rather, the argument is that individuals vary in how they would perform Type 2 activities, and there is no way of knowing in advance how a person would like to perform these activities.

5.2 The Logical Condition of Engagement

Nevertheless, in addition to epistemic challenges, there are also logical limitations which can restrict other entities from performing Type 2 activities on behalf of others.

According to the logical condition of engagement, activities which by definition require active subject participation cannot be delegated to autonomous systems. A defining feature of Type 2 activities is that what determines engagement in these activities is the individuals' quality of experience. What constitutes the quality of an experience varies for different people. However, for some individuals, the quality of an experience can be constituted by actively performing the activity by setting and performing the momentary goals. For example, for some people, the joyful experience of leisurely driving is not about where they go; it is in the very act of driving. When they are not driving a vehicle, they cannot engage in the activity of leisurely driving.

There are various activities in which it is necessary for one to actively perform the activity in order to be engaged in the activity. This necessity can have at least three different sources. In some cases, this necessity may be a personal preference. This is the case with the activity of leisurely driving. While some people require to actively drive a vehicle to take pleasure, some others may be happy for others to take them for a drive. In some other cases, this necessity may be imposed by authorities who determine the conditions of performing the activity. This is the case with the example of worshiping a deity which was given in the previous section. Whether one needs to personally perform certain actions in order to be engaged in the activity of worshiping is a matter which is determined by religious authorities. In some other activities, such as playing a board game, performing an improvised theatre, partying, moaning, respecting others or listening to music, this necessity may be in the logic of engagement in the activity. It is logically contradictory for a person to be engaged in these activities without personally performing them. A person cannot engage in the activity of playing a board game, performing an improvised theatre, partying, moaning, respecting others or listening to music by sitting aside and delegating the performance of these activities to others.

In any case, when being engaged in an activity requires actively performing the activity, the activity cannot be delegated to any other person or autonomous system. No entity, be it a human or an autonomous system, can perform such activities on behalf of others. As soon as these activities are delegated to other entities, they cease to be what they are. These activities, therefore, set a logical limit to human activities that autonomous systems can replace.

6 Conceptual Ramifications

Before discussing the ramifications of this analysis for the design and regulation of autonomous systems, a few words need to be said on its conceptual ramifications.

First, it is worth noting that the epistemic condition of performance and the logical condition of engagement divide activities into four types: those which satisfy

one condition but not the other, those which satisfy both and those which satisfy neither. For the brevity of analysis, Type 1 and Type 2 activities were chosen as two extreme cases where, respectively, both and neither of the conditions are satisfied. But there can be activities where the epistemic condition of performance can be met, but the logical condition of engagement cannot be met. Worshiping a deity can be an example here. Uttering certain words and using specific hand gestures to worship a deity can be taught to autonomous systems, but the engagement of the subject is a necessary condition for this activity. There can also be activities where the logical condition of engagement can be met, but the epistemic condition of performance can pose a challenge. Decorating a house according to a person's taste is an example of such activities. House decoration does not necessarily require subject engagement, but cannot be reliably taught to machines as it requires access to subjective knowledge and personal tastes.

Second, the distinctions made between different activities are not meant to be exclusive, as one person may be simultaneously engaged in two activities of different types by performing one set of actions. For example, a person may want to use a vehicle to go from one point to another, but they may also want to enjoy the drive. In this case, although the momentary goals must jointly contribute to the realization of the Type 1 activity of going from one place to another, they would not be purely instrumental, as in this case, the quality of the drive also matters. Because in such activities the way momentary objectives are set and followed is constitutive of some of the objectives of the activity, these activities fall outside the realm of activities which autonomous systems can perform.

Hence, although at the conceptual level the distinction between different types of activities remains clear, particular sets of actions cannot be universally classified into specific types of activities. As discussed in the previous section, it is not possible to decide the activity in which a person is engaged by observing the person's actions from a third person perspective. Some people may indeed take satisfaction from performing basic tasks, such as vacuum cleaning their home or weeding their garden, even though these tasks are often performed by most others solely as Type 1 activities to bring about specific worldly changes.

Autonomous systems can replace humans only in the performance of Type 1 activities where the epistemic condition of performance and the logical condition of engagement are both met. Type 1 activities are those which are performed solely to make changes to the states of affairs; they are activities in which objects and systems play objective instrumental roles.

Finally, autonomous systems are normal systems which have been 'modified' into autonomous systems. For example, autonomous vehicles are meant to be the modified versions of traditional cars. Non-autonomous systems can play different roles in different activities of humans. However, given the epistemic and logical limitations of system autonomy, when a system is turned autonomous, the system may not be able to be used in a variety of activities in which it used to play roles. When a system is redesigned as an autonomous system, it can only be employed in its capacity as a pure means to change the state of affairs. This conclusion, as discussed in the next section, has some ramifications on the design of autonomous systems.

7 Design Ramifications

Consider autonomous vehicles as an example here again. Some authors have argued that if in the future, autonomous vehicles prove to be safer than human drivers, then it should become illegal for humans to drive vehicles (Müller & Gogoll, 2020; Sparrow & Howard, 2017). Müller and Gogoll describe such a future with the phrase 'the angel car scenario', and Sparrow and Howard compare human drivers of that future to 'drunk robots'. Placing a ban on human driving in such a future would be to reduce the risks of human drivers endangering the lives of themselves and others due to being fatigued, anxious, fearful, distracted or under a drug influence. From a harm-reduction, consequentialist point of view, which is what Sparrow and Howard explicitly acknowledge to be the justification for their argument (Sparrow & Howard, 2017, p. 210), this is a valid point. If society's only concern is to make roads safer, then humans should not be allowed to drive vehicles when driving vehicles will result in more road casualties.

The argument to ban manual driving when autonomous vehicles become safer than humans is valid when Type 1 activities are concerned, which are practical problem-solving activities in which vehicles are employed as means to objective ends. This includes activities such as commuting or delivering goods. Activities like commuting or delivering goods pass the epistemic and logical conditions. In these activities, the performance of the activity is purely instrumental for the achievement of activity-dependent goals. What matters in these activities is the changes made to the outside world. There is no value in the act of driving when the only goal is simply to go from point A to point B. That is why in these activities we can follow the consequentialist argument and focus only on improving the outcomes. This means that we should not allow humans to drive, when robots are safer drivers: the users do not miss anything; and the roads become safer.

However, vehicles are not used only in Type 1 activities, such as taking the passengers from one place to another, where they play purely instrumental roles. Vehicles can also be used in activities where the performance of the activity has intrinsic values for their users, such as performing a burnout, racing or going for a leisurely drive. In such activities, the ethics of performing driving tasks cannot be reduced to consequentialist arguments which only consider the impacts of driving vehicles on society at large. Some values will go missing if users are deprived from performing these activities.

Therefore, there is more to consider in the ethical design and regulation of autonomous vehicles than making roads safer. The consequentialist policy of forbidding driving a car by humans clashes with our duty to respect user autonomy by allowing them to engage in activities which contribute to their personal identities. Autonomy is here defined as 'the authority to make decisions of practical importance to one's life' (Mackenzie, 2008, p. 512). Respecting others' autonomy is an important moral principle (Beauchamp & Childress, 2001; Gillon & Lloyd, 1994; Gillon, 2003), and is regarded as an important value in the design of autonomous systems (Verdiesen, 2017). In the context of designing and regulating the use of vehicles, respecting

others' autonomy requires allowing users to engage in various activities by using their vehicle.

What this means is that autonomous vehicles can operate as taxis, trams, delivery vehicles, bin collectors or in general, for transferring goods or passengers from a starting point to a destination. When vehicles are used as modes of transport, then their roles are only instrumental. These activities do not require user performance, and hence, user autonomy is not violated when they are deprived from actually driving vehicles. In these activities, the consequentialist argument of increasing road safety can be used to ban driving because no strong counter-argument stands against it. However, unless a user specifically requests otherwise, privately owned autonomous vehicles should always have the option for the user to switch to manual driving mode. This is a way for designers and regulators to respect users' autonomy.

This of course does not mean that user autonomy is the only consideration in the design of autonomous systems. Nor does it mean that user autonomy is a non-negotiable principle and overrides other moral and social considerations. However, user autonomy *is* a morally relevant factor, and there are those who actively practice their autonomy in the way they connect to their personal vehicles. For example, during the Summernats festival in Canberra, some people perform burnouts with their cars and show off their 'funky' vehicles. They dress up for the occasion and embrace a collective identity with others at the festival. In the case of Type 1 activities, such considerations are absent. But in activities where the quality of the user experience constitutes the activity, user autonomy becomes a morally relevant factor that needs to be considered next to other factors, such as public safety.

However, here it should also be noted that allowing humans to drive their vehicles does not necessarily open the door to avoidable road fatalities caused by careless driving or intentionally harmful acts. These risks can be mitigated by designing safety mechanisms into vehicles that would automatically trigger to prevent harmful outcomes. In the same way that some modern vehicles are equipped with autonomous emergency braking systems which stop the vehicle when it is about to hit an obstacle, autonomous vehicles can have further mechanisms which activate the brakes or take control over the steering wheel to prevent unnecessary harm. Such mechanisms can increase the safety of passengers and other road users by allowing 'AI-supervised human driving' (Müller & Gogoll, 2020). Through AI-supervised human driving, the outcome of reducing road fatalities can be achieved by autonomous safety mechanisms which trigger in circumstances when avoidable crashes are about to occur.

The conclusion made here about the design of autonomous vehicles can be generalized to other autonomous systems. Autonomous systems can operate in Type 1 activities where their role is purely instrumental. This includes activities such as changing a light bulb, vacuum cleaning a house, weeding a garden, deactivating a landmine, digging wells and painting objects on an assembly line. Systems that can be used in activities where they would play non-instrumental roles need to be designed with a manual control option.

8 Conclusion

According to activity realism, objects and systems cannot be identified or optimised without regard to the activities in which they are used. What should be sought to be optimised is the activities in which a system is used, and that may be achieved by making adjustments to the system. Each object or system can be used in different activities, and adjustments made to the object or system to optimise one activity may spoil others.

So, when we turn a system into an autonomous system, we need to ask ourselves: what is this thing that has become autonomous? In a way, it does not make sense to turn a system into an autonomous system in general. It is always particular uses of the system which can be performed autonomously by the system.

Now which particular use of systems can become autonomous? Only uses which satisfy the epistemic and the logical conditions. So, autonomous features can optimise technical systems only when these systems are used in activities which are purely defined by making changes to the outside world. In such activities, the means, methods and intermediate steps taken are only instrumental for the realisation of the goals of the activity. In the case of vehicles, such uses are limited to activities such as going from point A to point B or delivering goods. Other uses of vehicles cannot be performed autonomously by the vehicle.

This research does not respond to or replace the existing literature which address the limitations of system autonomy from moral, legal, technical and philosophical perspectives. It rather provides a new framework to be considered in the philosophical study and the design and regulation of autonomous systems.

Acknowledgement I am thankful for the comments that I received from Christine Boshuijzen-van Burken and the reviewers of this paper. I am also grateful to Sarah Greet for her comments and editing assistance.

Disclosure The author of this paper received funding from the Dutch Research Council (NWO) Platform for Responsible Innovation (NWO-MVI) as part of the DILEMA Project on Designing International Law and Ethics into Military Artificial Intelligence.

References

Aizawa, K. (2013). Introduction to "The material bases of cognition." *Minds & Machines, 23*, 277–286. https://doi.org/10.1007/s11023-013-9312-8

Beauchamp, T., & Childress, J. (2001). *Principles of biomedical ethics*. Oxford University Press.

Borgmann, A. (1984). *Technology and the character of contemporary life: A philosophical inquiry*. University of Chicago Press.
Coeckelbergh, M. (2021). Should we treat teddy bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds & Machines, 31*, 337–360.
Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309.
de Jong, R. (2019). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00120-4
Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese, 138*(3), 453–473.
Dreyfus, H. (1993). *What computers still can't do*. The MIT Press.
Gasser, G. (2021). The dawn of social robots: Anthropological and ethical issues. *Minds & Machines, 31*, 329–336. https://doi.org/10.1007/s11023-021-09572-9
Gillon, R. (2003). Ethics needs principles—four can encompass the rest—and respect for autonomy should be "first among equals." *Journal of Medical Ethics, 29*, 307–312.
Gillon, R., & Lloyd, A. (Eds.). (1994). *The principles of health ethics*. Wiley.
Haselager, P. (2013). Did I do that? Brain-computer interfacing and the sense of agency. *Minds & Machines, 23*, 405–418. https://doi.org/10.1007/s11023-012-9298-7
Krupiy, T. (2015). Of souls, spirits and ghosts: Transposing the application of the rules of targeting to lethal autonomous robots. *Melbourne Journal of International Law*, *16*(1), 145–202.
Mackenzie, C. (2008). Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy, 39*(4), 512–533.
Matthias, A. (2004). The responsibility gap in ascribing responsibility for the actions of automata. *Ethics and Information Technology, 6*, 175–183.
Müller, J., & Gogoll, J. (2020). Should manual driving be (eventually) outlawed? *Science and Engineering Ethics, 26*, 1549–1567. https://doi.org/10.1007/s11948-020-00190-9
Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*, 1201–1219.
Purves, D., Jenkins, R., & Strawser, B. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*, 851–872. https://doi.org/10.1007/s10677-015-9563-y
Roff, H. (2014). The strategic robot problem. *Journal of Military Ethics, 13*(3), 211–227.
Sartor, G., & Omicini, A. (2016). The autonomy of technological systems and responsibilities for their use. In: N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 39–74). Cambridge University Press. http://hdl.handle.net/1814/45234
Schulzke, M. (2011). Robots as weapons in just wars. *Philosophy & Technology, 24*, 293–306.
Searle, J. (1980). Minds, brains, and programs. In J. Haugeland (Ed.), *Minds design II: Philosophy, psychology, artificial intelligence* (pp. 183–204). MIT Press.
Sharkey, N. (2010). Saying 'No!' to Lethal Autonomous Targeting. *Journal of Military Ethics*, *9*(4), 369–383.
Soltanzadeh, S. (2019). A practically useful metaphysics of technology. *Techné: Research in Philosophy and Technology, 23*(2), 232–250. https://doi.org/10.5840/techne2019924103
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs, 30*(1), 93–116.
Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part c: Emerging Technologies, 80*, 206–215. https://doi.org/10.1016/j.trc.2017.04.014
Strawser, B. J. (2010). Moral predators: The duty to employ uninhabited aerial vehicles. *Journal of Military Ethics, 9*(4), 342–368.
Suchman, L. (1987). *Plans and situated action: The problem of human-machine communication*. Cambridge University Press.
Verdiesen, I. (2017). How do we ensure that we remain in control of our autonomous weapons? *AI Matters, 3*(3), 47–55.