Science and Engineering Ethics (2020) 26:2693–2708
https://doi.org/10.1007/s11948-020-00221-5

ORIGINAL RESEARCH/SCHOLARSHIP

Customizable Ethics Settings for Building Resilience and Narrowing the Responsibility Gap: Case Studies in the Socio-Ethical Engineering of Autonomous Systems

Sadjad Soltanzadeh[1] · **Jai Galliott**[1] · **Natalia Jevglevskaja**[1]

Received: 22 October 2019 / Accepted: 28 April 2020 / Published online: 26 May 2020

Abstract
Ethics settings allow for morally significant decisions made by humans to be programmed into autonomous machines, such as autonomous vehicles or autonomous weapons. Customizable ethics settings are a type of ethics setting in which the users of autonomous machines make such decisions. Here two arguments are provided in defence of customizable ethics settings. Firstly, by approaching ethics settings in the context of failure management, it is argued that customizable ethics settings are instrumentally and inherently valuable for building resilience into the larger socio-technical systems in which autonomous machines operate. Secondly, after defining the preliminary condition of responsibility attribution and demonstrating how ethics settings enable humans to exert control over the outcomes of morally significant incidents, it is shown that ethics settings narrow the responsibility gap.

Keywords Ethics settings · Autonomous machines · User autonomy · Resilience · Control · Responsibility gap

Introduction

The combination of a vehicle and its user can be studied as a small socio-technical system. Let us call this system a *user-vehicle system*.[1] User-vehicle systems are part of a larger system which can be called 'the network of transport and road use'. This larger system has other users as well, such as pedestrians and law enforcement authorities. The network of transport and road use operates successfully if its users are satisfied. Undesirable incidents are failures of the

[1] Here 'vehicle' is defined functionally, as a mobile technology with the capacity to carry passengers.

✉ Sadjad Soltanzadeh
s.soltanzadeh@unsw.edu.au

[1] School of Engineering and Information Technology, University of New South Wales, Canberra, Australia

network of transport and road use. Failure is defined broadly here, and includes anything from delays and accidents to air, noise, and light pollution.

Currently, the navigation of user-vehicle systems is performed by human users. However, many companies, such as Tesla, Alphabet, Audi, and BMW are investing in the production of level 4 or level 5 automated vehicles (CBINSIGHTS 2020). Level 4 or 5 automated vehicles can perform most or all navigational tasks in various road conditions without any input from their human users (SAE 2018). Although no level 4 or 5 autonomed vehicle has been introduced, different driver-assistance systems, such as autonomous braking systems and autonomous navigation in controlled environments have been successfully developed (CBINSIGHTS 2020). As more automated features are designed into vehicles, they will be able to navigate with little to no human input. Such vehicles are described by different terms, such as 'autonomous', 'self-driving', or 'driverless'. Here we use the phrase 'autonomous vehicle' as a general label to refer to level 4 or 5 vehicles.

Autonomous vehicles belong to the general category of products which are referred to as "autonomous machines". Autonomous machines are developed for other purposes, including military purposes. The Wave Glider, manufactured by Liquid Robotics, for example, is a wave and solar powered unmanned ocean robot. It serves as a communications gateway in a network of manned and unmanned assets, enabling seabed to space monitoring, and is suitable for anti-submarine warfare (ASW), intelligence, surveillance, and reconnaissance (ISR), as well as other military applications (Liquid Robotics n.d.).

The suitability of using the term 'autonomous' to refer to some technological products is of course questionable. This is particularly because current machines are not conscious or autonomous in the rich philosophical sense (Brey 2013; Nyholm 2018; Purves et al. 2015; Searle 1980). Nevertheless, this is an argument with which we shall not engage here. Our focus is on the *autonomy of users*, rather than the autonomy of the technology.

We discuss one particular way in which user autonomy can be supported in the design of autonomous machines; i.e., customizable ethics settings. Customizable ethics settings allow for morally significant decisions made by users to be programmed into autonomous machines.

"Failure Prevention and Failure Management" section introduces ethics settings in the context of failure management and building resilience into socio-technical systems. The section concludes by examining how the presence of ethics settings raises new moral questions, and that the more basic question to be asked is whether it is desirable to take further control over the outcomes of autonomous machines by implementing ethics settings into their design. "Understanding Ethics Settings" section argues for a positive answer to this question. The section examines the notion of "responsibility gap", and clarifies it in reference to the preliminary condition of responsibility attribution. The paper concludes by demonstrating how customizable ethics settings can function as a technological solution to narrow the responsibility gap and make the distribution of responsibilities more transparent.

Failure Prevention and Failure Management

Any socio-technical system is susceptible to failure. Overall, two general approaches can be adopted to address potential failures of socio-technical systems. The first approach is about *preventing* failure. This approach, which has led to different branches of Control Engineering (Åström and Murray 2008), requires identification of weak points, monitoring system parameters, and using reliable materials, data gathering techniques, and control mechanisms to prevent failure. The second approach is about *managing* and *(re-)directing* failure. Regardless of how strong the components and how advanced the data gathering and control mechanisms of a system are, some forms of failure may still occur. If all resources are put into preventing failure, unprevented failures can lead to further undesirable outcomes, create chaos, and become much more costly. As such, it is important to put as much thought into managing and redirecting failures as it is into preventing them. This second approach has led to the development of resilience building techniques in socio-technical systems, particularly in the sense of absorbing and responding to failures by retaining the core characteristics of the system (Doorn et al. 2018; Walker et al. 2004).

When a component of a system fails, its failure can cause other issues and potentially disable or create chaos in the whole system. To build resilience into the entire system, we need to decide how we want to contain failures so that the system can remain functional despite the failure of some of its components. This is achieved by identifying the essential features of the system and redirecting failures to protect those features. The essential features are what we value most and believe to be crucial in the desirable performance of the system.

Consider a socio-ecological system as an example. As this paper is being written, Australia is experiencing some of the most devastating and widespread bushfires in its history. As a result, the socio-ecological system has experienced many failures, such as the loss of more than half a billion plant and animal species, deaths of dozens of humans, and destruction of thousands of houses and buildings (British Broadcasting Corporation 2020). Making the system more resilient against bushfires requires identification *and* protection of those features of the system which are valued most and are believed to be crucial in the desirable performance of the socio-ecological system. So, it should first be asked: what are the most important features of the system? Is it human life? Animal and plant species? Government buildings? Telecommunication structures? Sport stadiums? Power plants? Secondly, depending on the answers given to the first question, mechanisms should be in place to protect the essential features of the system. Considering that identifying the mostly valued features of a system requires prioritization of values, building resilience requires inputs from social scientists, such as ethicists, sociologists, and lawyers.

Autonomous Vehicles and the Prevention and Management of Failure

Preventing failures of the network of transport and road use requires robust user-vehicle systems in terms of their hardware and software. It requires components that

do not easily break. It requires data gathering and control mechanisms which collect and process dynamic inputs from the environment. These inputs could be about the road and weather conditions, the speed limits and traffic lights, or the location, speed, and acceleration of other road users.

How can autonomous vehicles influence failure prevention? Autonomous vehicles can make the network of transport and road use more robust by reducing failure incidents. This can happen for at least two reasons. Firstly, as discussed by many authors (Gogoll and Müller 2017; Goodall 2014; Lin 2013; Sparrow and Howard 2017), autonomous vehicles should reduce the number of accidents caused by fatigue, intoxication, or distraction of drivers which account for many car crashes. Secondly, autonomous vehicles will be able to gather and process a wider range of environmental inputs compared to human drivers. A noteworthy example here is the ability of autonomous vehicles to communicate the 'intentions' of user-vehicle systems to each other. Human drivers get inputs about the intentions of other user-vehicle systems by noticing flashing indicators or brake lights. However, human drivers, even when they are most alert, are able to gather such information only from adjacent road users. On the contrary, autonomous vehicles are able to gather and process information not only from adjacent road users, but also from other road users whose navigation can affect their navigation. This can include gathering navigational data from every user-vehicle system on a high traffic road, including vehicles in blind spots or those which are further ahead on the road or are otherwise beyond the limits of human perception. Autonomous vehicles can gather and process the relevant data to make necessary adjustments to minimize the chance of collisions and other undesirable outcomes. These features of autonomous vehicles make user-vehicle systems and the entire network of transport and road use less susceptible to failure.

However, as noted above, not all attention should be given to failure prevention. No matter how robust a system is, some chance of failure remains. Think about driving on a road and facing a situation where the user-vehicle system needs to either steer the system to the left and potentially fall from a bridge, steer to the right and run over some cyclists, or go straight forward and hit pedestrians who have just stepped onto the road without properly scanning for oncoming vehicles. No matter what the system does, a failure, in the form of an undesirable outcome, will occur. Scenarios such as the one just described can be referred to as *trolley-problem-like scenarios*. Trolley-problem-like scenarios refer to situations where all available options lead to different forms of costly failures.[2]

What does building resilience into the network of transport and road use mean? Considering that each user-vehicle system is part of the broad network of transport and road-use, developing resilience means protecting the core characteristics of the broader network despite potential failures of user-vehicle systems. It requires identifying features and values which are central to the network of transport and road use and make it function in a desirable way. We may care about fairness, the protection of the environment, minimization of road casualties, or holding wrongdoers

[2] The reason why we call such scenarios "trolley-problem-like scenarios" is that they are generalized versions of the trolley problems (Foot 1967; Thomson 1985) which are widely used in different fields of applied and normative ethics.

accountable. By putting mechanisms in place which help us protect and promote these values despite failures of user-vehicle systems, we build resilience into the network of transport and road use. For instance, by managing road accidents in a way that that does not make us question our perception of justice and fairness in society, we make society more resilient towards road accidents.

Managing failures of user-vehicle systems is not always a straightforward task, because sometimes protecting and promoting one value implies sacrificing other values. As noted above, in the case of some fatal crashes, depending on the decisions which user-vehicle systems make, different parties will be killed. The trade-off between values occurs in the case of non-costly failures as well. For example, there is a trade-off between minimizing CO_2 emissions and taking a longer, more scenic route. In each instance of failure management, some values and valuables are protected and others are sacrificed.

Therefore, managing failure requires decision making mechanisms. Decision making mechanisms must be in place to decide what to protect and what to sacrifice. Then again, different parties may have incompatible ideas of what should be protected. Some people may value their own lives more than the lives of others, while others may value the opposite. Some may care about minimizing CO_2 emissions, and some others may be more interested in taking longer, more scenic routes. Thus, building resilience requires us to make decisions at another level as well. We need to determine who should have the authority to influence failure management by prioritizing values and valuables.

Understanding Ethics Settings

Ethics Settings as a Technological Means for Failure Management

Traditionally, the immediate management of the failures of user-vehicle systems has been the sole responsibility of users, because users have been the only social group who have had control over the navigation of user-vehicle systems. However, the navigation of autonomous vehicles is controlled by algorithms, and unless specific commands are programmed into the vehicle, the decision regarding 'what values to protect?' will be made by algorithms not equipped for the purpose.

This is where ethics settings come into play. Ethics settings are settings that can be programmed into autonomous vehicles to manage different forms of failure by making autonomous vehicles follow human decisions. Technically, ethics settings are higher-order decision-making commands, such as 'take the scenic route' or 'protect the passengers'. In the absence of ethics settings, the user-vehicle system follows driving decisions that machine learning algorithms make.[3] Decisions made by such

[3] Currently, there are three main types of machine learning algorithms, namely, reinforcement learning, supervised, and unsupervised machine learning algorithms. These algorithms are dynamic in the sense that they do not follow rigid input/output relationships. They can readjust by each new input or 'experience' that the machine undergoes. This allows machines to constantly improve their function. For instance, an autonomous vehicle may constantly adjust itself to find an optimal compromise between keeping a safe distance from adjacent vehicles, avoiding unnecessary brakes, and driving smoothly and

algorithms are not always predictable or transparent (Ananny and Crawford 2016; Pasquale 2015), and may appear as random to human observers. By implementing ethics settings, humans can program their intentions into autonomous systems instead of subjecting themselves to unpredictable decisions made by algorithms. For example, users may adjust the settings so that the vehicle always gives more weight to the protection of the passengers in car crashes.

Ethics settings may also be used to manage less costly, yet still morally relevant failures. Consider a setting which requires the vehicle to choose a route which would minimize its greenhouse gas emissions, a setting which stops the vehicle to let an animal cross the road, or a setting which slows down to allow another vehicle to join the traffic on a busy road. Such settings influence the navigation of the user-vehicle system in the routes chosen by the autonomous vehicle, its driving style, and its speed and acceleration at each point in time. Though morally significant, these settings are not as significant as a setting that can be used to manage immediate life or death scenarios.[4]

As such, ethics settings can be seen as a method of failure management and building resilience into the network of transport and road use. By programming their preferences into autonomous vehicles, humans can manage failures by protecting their core values.

Types of Ethics Settings

The idea of ethics settings can be realized in at least two different ways. Firstly, ethics settings can be pre-programmed into autonomous vehicles by manufacturers. This could be achieved in consultation with ethics committees or regulatory bodies. Gogoll and Müller refer to this type of ethics setting as the 'mandatory ethics setting' (Gogoll and Müller 2017). Such settings are mandatory in the sense that users do not have the freedom to change the settings. All user-vehicle systems are mandated to follow the same instructions when they face similar scenarios. For instance, they may all aim to maximize fuel efficiency, or they may all aim to minimize the number of casualties in trolley-problem-like scenarios.

A second type of ethics settings can be referred to as "personal ethics settings" (Gogoll and Müller 2017) or "customizable" or "adjustable ethics settings" (Lin 2014a). Customizable ethics settings are a type of ethics setting in which morally significant decisions are made by the users of autonomous vehicles. Ideally, these settings should be readjustable on each trip. For example, a user who cares for the

Footnote 3 (continued)
time-efficiently. Each trip can provide a new experience for the autonomous vehicle to readjust its driving style to find that optimal zone.

Ethics settings are not meant to be learned and dynamically readjusted by machines, and they do not fall into the category of machine learning algorithms. Ethics settings are meant to be incorporated into decision making mechanisms as higher-order commands. The decision making mechanisms of autonomous machines constantly considers the settings adjusted by users before executing any decision.

[4] Millar refers to settings which do not involve management of costly failures as 'low-stakes ethics settings' and contrasts them with 'high-stakes ethics settings', which occur in the form of trolley-problem-like scenarios (Millar 2017).

environment may customize the settings to minimize greenhouse gas emissions. However, if 1 day they want to use the vehicle to quickly reach a certain destination, they can readjust the vehicle to follow the quickest route on that occasion.

Other ways of implementing ethics settings can be achieved through a combination of personal and mandatory ethics settings where some of the settings are hard-programmed by the manufacturers and others can be readjusted by users. Here regulators can step in and draw the line with respect to what users should be allowed to adjust by means of ethics settings. For example, regulators can prohibit manufacturers from incorporating a customizable setting into autonomous vehicles by means of which users can discriminate between other road users based on their race or gender. Such a regulation would be similar to requiring car manufacturing companies to produce vehicles with functioning seat belts, and requiring drivers and passengers to fasten their seat belts when using their vehicles. Although these regulations restrict users' freedom, minor restrictions which are aimed at protecting others from severe harms are justified. It is justified to make users slightly uncomfortable by mandating the use of seat belts which can save lives during car crashes. It is justified to exclude settings by means of which the users of autonomous vehicles can act upon their racist, sexist, or similarly harmful intentions.[5]

The Instrumental and Inherent Values of Customizable Ethics Settings

Regardless of who makes the decision, decisions made to manage and redirect failures are inherently morally significant for at least two reasons. Firstly, because they deal with harms inflicted on humans, and any situation which involves managing harms inflicted on humans is morally significant.

Secondly, ethics settings are inherently morally significant also because they pertain to one of the most fundamental notions of moral philosophy and moral psychology, namely, autonomy. Autonomy is important as it provides a basis for bearing rights and responsibilities. We exercise our autonomy by making decisions that affect our lives. When we are deprived of the possibility of making decisions which affect our lives, our autonomy is compromised.

[5] Such restrictions are justified particularly if we uphold a substantive relational notion of autonomy, according to which respecting users' autonomy does not mean acceding to their requests, regardless of the content of their requests. Respecting autonomy, rather, involves "an obligation to promote autonomy" (Mackenzie 2008, p. 514). Culturally oppressive ideas can impair an individual's autonomy. In such cases, promoting autonomy requires understanding and removing oppressive ideas and unjustified biases which influence one's motivations, such as racist and sexist ideas.

However, even if we uphold the individualistic notion of autonomy (also known as the procedural notion of autonomy), according to which, regardless of their contents, an individual's subjectively scrutinized decisions should be respected, we can still justify minimal restrictions to limit what users can adjust through ethics settings. Manufacturers and regulators have an obligation to respect users' autonomy by allowing them to control the outcome of all morally significant decisions. However, they also have an obligation to protect others from harm or at least not facilitate harmful behaviours. We believe that the latter obligation outweighs the former. It is wrong to facilitate harmful behaviours by implementing ethics settings by means of which users can wrongfully discriminate between other users.

Therefore, both types of ethics settings are instrumentally valuable for resilience building because they allow humans to manage failures by prioritizing values and programming their priorities into autonomous machines. In addition to this instrumental value, customizable ethics settings also intrinsically embody building resilience into the network of transport and road use, regardless of how ethics settings are adjusted. This is because customizable ethics settings support and respect user autonomy which is itself an intrinsic value. Autonomy, "the authority to make decisions of practical importance to one's life" (Mackenzie 2008, p. 512), is intrinsically valuable as it is used to define the notion of personhood and justify the value given to a person's life (Walker and Lovat 2015; Warren 2000). Respect for autonomy is one of the fundamental principles of applied ethics (Beauchamp and Childress 2001; Gillon and Lloyd 1994), and according to some, it is the most important ethical principle (Gillon 2003). However, autonomy is conditioned by freedom to act. When someone is deprived from making choices pertaining to their lives and personal identity, their autonomy is restricted. Customizable ethics settings work as technical means that allow users to make choices in the way in which they want to manage their vehicle. Hence, customizable ethics settings make the network of transport and road use more resilient because even in the face of failures, one of the core social values is protected.[6]

Post-phenomenology of Ethics Settings

Through the example of the technology of obstetric ultrasound imaging, Verbeek (2008, 2011) demonstrated how the use of a particular technology can open up new dimensions of moral thinking. With the availability of obstetric ultrasound imaging, parents now face at least two morally significant questions. First, they need to decide whether or not they want to use this technology to screen the foetus' development. And secondly, they need to decide whether and how they want to proceed with pregnancy after receiving the ultrasound results (Verbeek 2008, 2011). The important point here is that in the absence of obstetric ultrasound imaging, these ethical questions would not arise. Prior to the invention of this technology, it was not feasible for humans to make choices, intervene, and manage the outcome of pregnancy by studying dynamic images of unborn foetuses.

The moral impact of ethics settings in the context of using autonomous vehicles is similar to the moral impact of obstetric ultrasound imaging in the context of pregnancy. They both open possibilities of managing potentially costly failures. Their sheer presence requires us to make morally significant decisions. Ethical

[6] The way in which the idea of ethics settings was initially introduced here followed a utilitarian approach in relation to the consequences of the settings on the whole network of transport and road use. However, other ethical theories can also be used to justify ethics settings. Respect for autonomy is a deontological principle which is used to explain why customizable ethics settings are inherently valuable. Moreover, users who are motivated by fostering and exhibiting virtuous traits in an Aristotelian framework can adjust the settings so that their autonomous vehicle makes choices which are in line with their desired virtues. For instance, altruistic users can adjust the settings so that their vehicles slow down to allow other vehicles to join the traffic on busy roads.

questions which arise as a result of implementing ethics settings are 'What is the morally acceptable decision in each instance of failure management?' and 'Who should make the decision?' In the absence of ethics settings, these questions would be practically irrelevant. They would be practically irrelevant because, regardless of whether we find clear answers to them, we would not have had the technological means to control the outcome.

Lack of control over the outcome of some morally significant scenarios is what we witness with non-autonomous vehicles. Currently, drivers make a split-second 'reaction' if they face trolley-problem-like scenarios. There simply is not enough time for them to recognize and reflect on morally relevant factors to make a justified decision. Autonomous vehicles that do not have any form of ethics settings, too, deprive their users from exerting control over the outcome of costly failures. Such autonomous vehicles would follow commands generated by algorithms which are not transparent and may be perceived as random. Ethics settings turn reactions and randomness into decisions and deliberate actions (Lin 2014b). By using ethics settings, we can reflect on scenarios, decide which outcomes we would want to bring about, and program them into the autonomous vehicle.

What this means for the design and evaluation of autonomous vehicles is that before addressing either of the two moral questions in relation to the failure management of user-vehicle systems (i.e., 'Who should make morally significant decisions?' and 'What is the morally acceptable decision in each case?'), we need to ask a more basic question. This more basic question is whether we want to have ethics settings altogether.

Ethics Settings as a Technological Solution to Narrow the Responsibility Gap

In addition to supporting user autonomy and building resilience into socio-technical systems, ethics settings can also narrow the responsibility gap.[7]

There are debates over whether autonomous machines open up a responsibility gap. It has been argued that when autonomous machines make decisions which are unpredictable and uncontrollable by humans, no one can be held responsible for machines' "wrongdoings", and therefore, there will be a responsibility gap (Matthias 2004; Sparrow 2007, 2016; Roff 2013, 2014) or a retribution gap (Danaher 2016; de Jong 2019) associated with their function. In response, it has been argued that the actions of some autonomous machines, particularly autonomous weapons, are constrained by the hierarchical structure of the military and the (implicit) agreements between the State, its citizens, and its military force. Hence, existing social structures fill the alleged responsibility gap (Galliott 2015; Leveringhaus 2016; Schulzke 2013). Some others argue that although autonomous weapons do not create a responsibility gap, they do create a related gap, namely, a 'blameworthiness' gap (Simpson and Müller 2016).

[7] Here our focus is on moral responsibility. For the sake of brevity, we only use the term responsibility.

Our focus here is on the impacts of ethics settings on the distribution of responsibilities. We first provide a working definition of "responsibility gap" and the conditions in which a responsibility gap can occur, and then we argue that ethics settings can function as a technological solution to narrow the responsibility gap.

The Responsibility Gap and the Preliminary Condition of Responsibility Attribution

What is meant by a responsibility gap? A responsibility gap occurs when there is a gap in the attribution of responsibility in situations when one can reasonably search for a responsible party. For a responsibility gap to occur, two conditions need to be met. The first condition is what we call "the preliminary condition of responsibility attribution". The preliminary condition of responsibility attribution determines the situations in which we can reasonably look for responsible parties. This condition does not specify the responsible party. It rather validates the very search for the responsible party. The second condition stipulates that for a responsibility gap to occur, no one should be able to bear responsibility.

In what situations is the preliminary condition of responsibility attribution met? The preliminary condition of responsibility attribution is met when we face morally significant outcomes of a controllable incident. In other words, morally significant outcomes of controllable incidents warrant the search for responsible parties. Here it is important to clarify the terms and phrases used in this definition.

The first phrase to clarify is "morally significant". It is important for an incident to be morally significant (e.g., a person being harmed) for us to care about the attribution of responsibility. For morally insignificant incidents (e.g., eating an apple and a banana for breakfast), no attribution of moral responsibility is required.

The second term to clarify is "incident". Let us explain why it is important to focus on incidents rather than actions. Incidents constitute a broader group of happenings than actions. An action needs to be intended whereas incidents may or may not be intentional. Although all entities bear responsibility for their intentional actions, they can also be held responsible for the outcomes of non-intended incidents. Think about holding people responsible for negligence. When a morally significant incident can be prevented, the parties who have failed to act to prevent the incident can be held responsible for the outcome. This is despite the fact that the outcome was not brought about as a result of their intentional actions.

However, it is not the case that the morally significant outcomes of *all* incidents warrant the search for responsible parties. We can reasonably search for responsible parties only when we face morally significant outcomes of *controllable* incidents. We cannot justifiably look for responsible parties when we face morally significant, negative outcomes of natural incidents, such as lightning. This is because most natural incidents are not currently controllable. However, when we are harmed as a result of others' actions or negligence, we can reasonably search for responsible parties. This is because we believe the harm could have been avoided.

Hence, as others have noted, control is an important condition of responsibility attribution (Nelkin 2013; Sand 2019; Sparrow 2007; Zimmerman 2002). This

condition is often expressed in the Control Principle: "We are morally assessable only to the extent that what we are assessed for depends on factors under our control" (Nelkin 2013). What distinguishes our focus is that while others often use the Control Principle to discuss whether *a particular party* is responsible for a particular incident, here we use controllability to discuss whether *a particular incident* is one for which *anyone* can be held responsible. If an incident is not controllable, then we cannot look for any responsible individual. This can be understood as a generalized version of the Control Principle.

A responsibility gap occurs when despite the fact that the preliminary condition of responsibility attribution is met, no one can be held responsible. It occurs when no specific party can be held responsible for morally significant outcomes of a controllable incident.

So what are the concrete cases of responsibility gaps? Although the responsibility gap is often discussed in the context of autonomous machines and in particular, lethal autonomous weapons, not all instances of responsibility gaps are limited to the implementation of autonomous machines. Think about a society in which it is agreed that due to the very low likelihood of thunderstorms in the region and the very high costs of lightning rods, building companies are not required to install lightning rods on buildings. If in that society lightning does strike and kills a number of people in a building which does not have a lightning rod, a controllable incident has brought about a morally significant outcome without anyone being responsible for it. The outcome could be avoided by installing a lightning rod. But considering that the building company was not required to install a lightning rod, they cannot be held responsible for the outcome. The decision makers cannot be responsible because the decision to not require lightning rods for buildings was justified by the available evidence on the risks of lightning in the region. The lightning rod manufacturers cannot be held responsible either because the deaths were not caused by a malfunctioning rod.

A responsibility gap can occur in the operation of autonomous machines when a controllable incident related to their operation generates morally significant outcomes, yet no party can justifiably bear the responsibility for the outcome. For example, if society, through a democratic process, decides to allow the production of autonomous machines without requiring manufacturers to consider the moral implications of these machines in their design, then we are likely to face responsibility gaps. It is likely that autonomous machines produce outputs which are morally significant. Foreseeable morally significant incidents brought about by autonomous machines are controllable. This is because programmers can be required to make adjustments in how they program autonomous machines in order to manage morally significant outcomes. But if programmers are not required to do so, they cannot be held responsible. Users cannot be held responsible either because they have no control over the incidents caused by autonomous machines. Here we face a situation where no one can be held responsible for morally significant outcomes of a controllable incident.

When autonomous machines are introduced into society without their moral implications being considered in the process of design, unintended harms caused by autonomous systems resemble harms caused by natural disasters. In either case,

no one can be held responsible for the morally significant negative outcomes. In this way, integrating unpredictable autonomous machines into society can be seen as creating possibilities of more environmental disasters. People may get trapped under debris after an earthquake, get struck by lightning, die of tetanus, get bitten by a snake, get driven over by an autonomous vehicle, or get shot by an autonomous killer robot.

However, although their occurrence is to a great extent uncontrollable, we put mechanisms in place to manage harms caused by natural disasters. We may not be able to stop lightning from happening, but we can install lightning rods on tall buildings. We may not be able to eradicate or cure tetanus, but we can use vaccination to immunize ourselves against the bacteria. We may not be able to stop earthquakes, but we devise safety standards for buildings and require engineers to follow them.

Safety mechanisms, of course, need not only be used to manage the outcomes of natural events. Safety mechanisms can be used to manage the outcomes of incidents related to new technologies. New technologies can often pose safety concerns which remain uncontrollable unless further technologies are developed to manage the risks associated with them. Bikes pose a risk to riders in that if they fall, they may incur a serious head injury. Prior to the invention of bike helmets, the risk of incurring a head injury while coming off the bike was uncontrollable in the same way that prior to the invention of lightning rods, the risk of getting hurt when lightning strikes was uncontrollable. However, the introduction of bike helmets has made this risk controllable.

Ethics Settings and the Attribution of Responsibility

New technologies often pose certain risks which make us cautious about introducing them into society. But adding further features to the technologies can make some of the potential risks more controllable. When specific features are introduced to control the risks, relevant social groups can gain new responsibilities. These responsibilities are defined in relation to the roles that people acquire in the context of using the features for risk management.

Thus, mechanisms used to manage unintended incidents not only help us to control undesirable outcomes, they also open avenues for the ascription of responsibility. The availability of the technology of bike helmets, for example, has led society to generate protocols which riders and helmet manufacturers are expected to follow. Manufacturers are now responsible for producing safe bike helmets, and riders are responsible for wearing helmets while riding their bikes. The availability of the technology of obstetric ultrasound imaging, as noted earlier, has made doctors responsible to inform expecting parents of the risks of pregnancy and parents responsible for the choices that they make in response to ultrasound results (Verbeek 2008, 2011).

Ethics settings are specific features which, similar to other technologies, have problem solving properties defined in relation to the context in which they are used (Hickman 2001; Soltanzadeh 2015, 2016). One of the perceived risks of autonomous vehicles, and autonomous machines in general, is the potential creation of responsibility gaps. The fact that autonomous machines might open up responsibility gaps

has often been used to advocate against the development and use of these machines. Banning the use of autonomous machines is of course one way to solve the problem of the responsibility gap. However, ethics settings provide another solution to this problem without depriving us from benefits which these machines can provide. Ethics settings can narrow the responsibility gap because they allow control over a range of morally significant outcomes. Relevant social groups acquire new responsibilities according to their roles in relation to ethics settings. Manufacturers will be responsible to integrate ethics settings into the design of autonomous vehicles, and users will be responsible for the customization of the ethics settings.

In the context of autonomous weapons, too, ethics settings help to narrow the responsibility gap which has been a concern in the debate over the permissibility of the development and use of such weapons. The availability and use of ethics settings means more decisions are intended by humans (in this case, the responsible authority in the chain of command), and fewer morally significant outcomes result from the operation of potentially unpredictable algorithms. Manufacturers will be responsible to include customizable ethics settings in the design of autonomous weapons, and operators and/or commanders will be responsible for the intended effects of their preferred settings. Ethics settings enable humans to exert more control over the outcomes of weapon use, which in turn will make the distribution of responsibilities more transparent.

Therefore, although currently ethics settings have been discussed solely in the context of autonomous vehicles, the design of other autonomous machines can also benefit from these settings. In fact, the idea of ethics settings for weapons (and their support) systems deployed in combat may even be more profound where law is either silent on the issue or is open to diverse (possibly conflicting) interpretations. Fundamental principles of the law of armed conflict permit status-based attacks against adversary with combat power highly likely to cause death unless and until the adversary is rendered physically incapable of participating in combat (Corn et al. 2013). While law permits employing methods and means of warfare which are likely to produce death as a first resort, ethical norms may require a soldier to use less harmful means. For example, if an adversary can be incapacitated with no additional risk to State's own or allied forces, then any harm inflicted in excess of that would be legally acceptable, but ethically impermissible. Customizable ethics settings enable militaries to program these additional moral concerns into their autonomous weapons. Such settings would promote decision-making transparency and clarify the responsibility distribution.

Conclusion

Autonomous machines are likely to make the broad socio-technical systems in which they operate safer by minimizing the failure incidents caused by human error. However, mechanisms still need to be in place to make socio-technical systems more resilient, so that failures of autonomous machines can be contained and managed.

Two main arguments in defence of customizable ethics settings were provided here. Firstly, customizable ethics settings are instrumentally and inherently valuable

for failure management and building resilience into socio-technical systems, such as the network of transport and road use. Different types of ethics settings are instrumentally valuable for building resilience into socio-technical systems because they enable humans to manage failures by programming their values into autonomous machines. Customizable ethics settings also intrinsically embody building resilience into socio-technical systems, because they always protect and respect user autonomy which is itself an intrinsic value regardless of how the settings are adjusted.

Secondly, ethics settings can be used as a technological solution to narrow the responsibility gap. New technological features can enable us to exert control over the outcomes of previously uncontrollable events. If we do not implement these features, controllable incidents may have morally significant outcomes for which no one might bear responsibility. Customizable ethics settings will hold manufacturers responsible for incorporating these settings into the design of autonomous machines, and will hold users responsible for the choices that they make.

Acknowledgements We are thankful to the anonymous reviewers whose constructive feedback helped to improve the quality of this paper.

Funding This study was supported by Air Force Office of Scientific Research (Grant No. FA9550-18-1-0181).

References

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645.

Åström, K. J., & Murray, R. (2008). *Feedback systems: An introduction for scientists and engineers*. Oxfordshire: Princeton University Press.

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

Brey, P. (2013). From moral agents to moral factors: The structural ethics approach. In P. Kroes & P.-P. Verbeerk (Eds.), *The moral status of artifacts* (pp. 125–142). Berlin: Springer.

British Broadcasting Corporation (BBC). (2020). *Australia fires: A visual guide to the bushfire crisis.* Retrieved January 31, 2020 from https://www.bbc.com/news/world-australia-50951043.

CBINSIGHTS. (2020). *40+ Corporations working on autonomous vehicles.* Retrieved March 4, 2020 from https://www.cbinsights.com/research/autonomous-driverless-vehicles-corporations-list/.

Corn, G., Blank, L., Jenks, C., & Jensen, E. T. (2013). Belligerent targeting and the invalidity of a least harmful means rule. *International Law Studies, 89*, 536–626.

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

de Jong, R. (2019). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics, 26*(8), 727–735. https://doi.org/10.1007/s11948-019-00120-4.

Doorn, N., Paolo, G., & Colleen, M. (2018). A multidisciplinary definition and evaluation of resilience: The role of social justice in defining resilience. *Sustainable and Resilient Infrastructure, 4*(3), 112–123. https://doi.org/10.1080/23789689.2018.1428162.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15. https://doi.org/10.1093/0199252866.001.0001.

Galliott, J. (2015). *Military robots: Mapping the moral landscape*. Brookfield: Ashgate Publishing.

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*. Chichester: Wiley.

Gogoll, J., & Müller, J. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*(3), 681–700. https://doi.org/10.1007/s11948-016-9806-x.

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*, 58–65. https://doi.org/10.3141/2424-07.

Hickman, L. (2001). *Philosophical tools for technological culture: Putting pragmatism to work*. Bloomington: Indiana University Press.

Leveringhaus, A. (2016). *Ethics and autonomous weapons*. London: Palgrave Pivot.

Lin, P. (2013). *The ethics of saving lives with autonomous cars are far murkier than you think*. Retrieved from https://www.wired.com/opinion/2013/07/the-surprising-ethics-of-robot-cars. Accessed 10 June 2019.

Lin, P. (2014a). *Here's a terrible idea: Robot cars with adjustable ethics settings*. Retrieved from https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/. Accessed 10 June 2019.

Lin, P. (2014b). *The robot car of tomorrow may just be programmed to hit you*. Retrieved from https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/. Accessed 10 June 2019.

Liquid Robotics. (n.d.). *Reimagine ocean monitoring and operations. unmanned robots powered by nature*. Retrieved from https://www.liquid-robotics.com/wave-glider/overview/.

Mackenzie, C. (2008). Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy, 39*(4), 512–533. https://doi.org/10.1111/j.1467-9833.2008.00440.x.

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

Millar, J. (2017). Ethics settings for autonomous vehicles. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190652951.001.0001.

Nelkin, D. (2013). Moral luck. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Winter 2013 edition)*. https://plato.stanford.edu/archives/win2013/entries/moral-luck/. Accessed 1 Nov 2019.

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap II. *Philosophy Compass*. https://doi.org/10.1111/phc3.12506.

Pasquale, F. (2015). *The Black Box Society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Purves, D., Jenkins, R., & Strawser, B. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y.

Roff, H. (2013). Killing in war: Responsibility, liability and lethal autonomous robots. In F. Allhoff, N. Evans, & A. Henschke (Eds.), *Routledge handbook of ethics and war: Just war theory in the 21st century*. London: Routledge.

Roff, H. (2014). The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics, 13*(3), 211–227. https://doi.org/10.1080/15027570.2014.975010.

SAE (Society of Automotive Engineers). (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Retrieved from https://www.sae.org/standards/content/j3016_201806/. Accessed 10 June 2019.

Sand, M. (2019). Did Alexander Fleming deserve the nobel prize? *Science and Engineering Ethics, 26*, 899–919. https://doi.org/10.1007/s11948-019-00149-5.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*, 203–219. https://doi.org/10.1007/s13347-012-0089-0.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756.

Simpson, T., & Müller, V. (2016). Just war and robots' killings. *The Philosophical Quarterly, 66*(263), 302–322. https://doi.org/10.1093/pq/pqv075.

Soltanzadeh, S. (2015). Humanist and nonhumanist aspects of technologies as problem solving physical instruments. *Philosophy & Technology, 28*(1), 139–156. https://doi.org/10.1007/s13347-013-0145-4.

Soltanzadeh, S. (2016). Questioning two assumptions in the metaphysics of technological objects. *Philosophy & Technology, 29*(2), 127–135. https://doi.org/10.1007/s13347-015-0198-7.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Sparrow, R. (2016). Robot 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.

Thomson, J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.

Verbeek, P.-P. (2008). Obstetric ultrasound and the technological mediation of morality: A postphenomenological analysis. *Human Studies, 31*(1), 11–26. https://doi.org/10.1007/s10746-007-9079-0.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: Chicago University Press.

Walker, B., Holling, C., Carpenter, S., & Kinzig, A. (2004). Resilience, adaptability and transformability in social-ecological systems. *Ecology and Society, 9*(2), 5–13. https://doi.org/10.5751/ES-00650-090205.

Walker, P., & Lovat, T. (2015). Concepts of personhood and autonomy as they apply to end-of-life decisions in intensive care. *Medicine, Health Care, and Philosophy, 18*(3), 309–315. https://doi.org/10.1007/s11019-014-9604-7.

Warren, M. A. (2000). *Moral status: Obligations to persons and other living things*. New York: Oxford University Press.

Zimmerman, M. (2002). Taking luck seriously. *Journal of Philosophy, 99*(11), 553–576. https://doi.org/10.2307/3655750.