Implementable Ethics for Autonomous Vehicles

5

J. Christian Gerdes, Sarah M. Thornton

Content

J. C. Gerdes (✉)
Stanford University, Dept. of Mechanical Engineering, Center for Automotive Research at Stanford, USA
gerdes@stanford.edu/gerdes@cdr.stanford.edu

S. M. Thornton
Stanford University, Dept. of Mechanical Engineering, USA
smthorn@stanford.edu

M. Maurer et al. (Hrsg.), *Autonomes Fahren*,
DOI 10.1007/978-3-662-45854-9_5,

5.1 Introduction

As agents moving through an environment that includes a range of other road users – from pedestrians and cyclists to other human or automated drivers – automated vehicles continuously interact with the humans around them. The nature of these interactions is a result of the programming in the vehicle and the priorities placed there by the programmers. Just as human drivers display a range of driving styles and preferences, automated vehicles represent a broad canvas on which the designers can craft the response to different driving scenarios. These scenarios can be dramatic, such as plotting a trajectory in a dilemma situation when an accident is unavoidable, or more routine, such as determining a proper following distance from the vehicle ahead or deciding how much space to give a pedestrian standing at the corner. In all cases, however, the behavior of the vehicle and its control algorithms will ultimately be judged not by statistics or test track performance but by the standards and ethics of the society in which they operate.

In the literature on robot ethics, it remains arguable whether artificial agents without free will can truly exhibit moral behavior [1]. However, it seems certain that other road users and society will interpret the actions of automated vehicles and the priorities placed by their programmers through an ethical lens. Whether in a court of law or the court of public opinion, the control algorithms that determine the actions of automated vehicles will be subject to close scrutiny after the fact if they result in injury or damage. In a less dramatic, if no less important, manner, the way these vehicles move through the social interactions that define traffic on a daily basis will strongly influence their societal acceptance. This places a considerable responsibility on the programmers of automated vehicles to ensure their control algorithms collectively produce actions that are legally and ethically acceptable to humans.

An obvious question then arises: can automated vehicles be designed a priori to embody not only the laws but also the ethical principles of the society in which they operate? In particular, can ethical frameworks and rules derived for human behavior be implemented as control algorithms in automated vehicles? The goal of this chapter is to identify a path through which ethical considerations such as those outlined by Lin, Bekey and Abney [2] and Goodall [3] from a philosophical perspective can be mapped all the way to appropriate choices of steering, braking and acceleration of an automated vehicle. Perhaps surprisingly, the translation between philosophical constructs and concepts and their mathematical equivalents in control theory proves to be straightforward. Very direct analogies can be drawn between the frameworks of consequentialism and deontological ethics in philosophy and the use of cost functions or constraints in optimal control theory. These analogies enable ethical principles that can be described as a cost or a rule to be implemented in a control algorithm alongside other objectives. The challenge then becomes determining which principles are best described as a comparative weighting of costs from a consequentialist perspective and which form the more absolute rules of deontological ethics.

Examining this question from the mathematical perspective of deriving control laws for a vehicle leads to the conclusion that no single ethical framework appears sufficient. This

echoes the challenges raised from a philosophical perspective by Wallach and Allen [4], Lin et al. [2] and Goodall [3]. This chapter begins with a brief introduction to principles of optimal control and how ethical considerations map mathematically into costs or constraints. The following sections discuss particular ethical reasoning relevant to automated vehicles and whether these decisions are best formulated as costs or constraints. The choice depends on a number of factors including the desire to weigh ethical implications against other priorities and the information available to the vehicle in making the decision. Since the vehicle must rely on limited and uncertain information, it may be more reasonable for the vehicle to focus on avoiding collisions rather than attempting to determine the outcome of those collisions or the resulting injury to humans. The chapter concludes with examples of ethical constraints implemented as control laws and a reflection on whether human override and the ubiquitous "big red button" are consistent with an ethical automated vehicle.

5.2 Control Systems and Optimal Control

Chapter 4 outlined some of the ethical frameworks applicable to automated vehicles. The first step towards implementing these as control algorithms in a vehicle is to similarly characterize the vehicle control problem in a general way. Figure. 5.1 illustrates a canonical schematic representation of a closed-loop control system. The system consists of a plant, or object to be controlled (in this case, an autonomous vehicle), a controller and a set of goals or objectives to satisfy. The basic objective of control system design is to choose a set of control inputs (brake, throttle, steering and gear position for a car) that will achieve the desired goals. The resulting control laws, in general, consist of a priori knowledge of the goals and a model of the vehicle (feedforward control) together with the means to correct errors by comparing measurements of the environment and the actual vehicle motion (feedback control).

Many approaches have been formulated over the years to produce control laws for different goals and different types of systems. One such method is optimal control, originally developed for the control of rockets in seminal papers by Pontryagin and his colleagues [5].

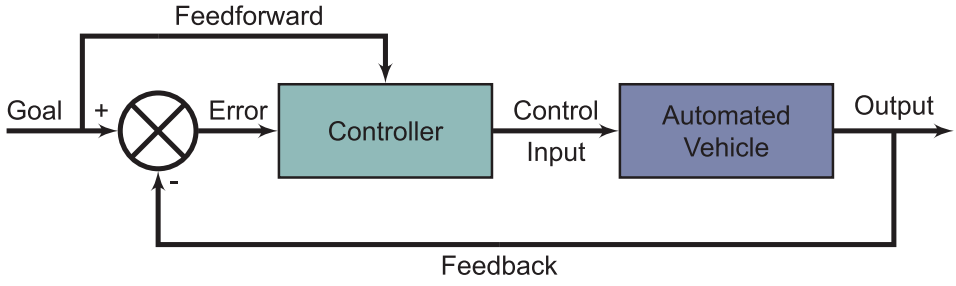

Fig. 5.1 A schematic representation, or block diagram, of a control system showing how control inputs derive from goals and feedback

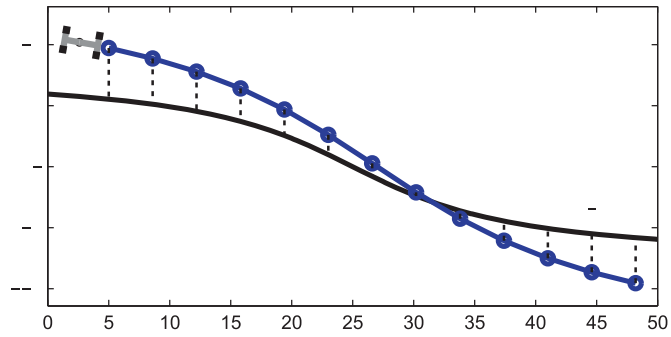

Fig. 5.2 Generating a cost from the difference between a desired path (black) and the vehicle's actual path (blue)

In a classic optimal control problem, the goal of the system is expressed in the form of a cost function that the controller should seek to maximize or minimize. For instance, the goal of steering a vehicle to a desired path can be described as minimizing the error between the path taken by the vehicle and the desired path over a certain time horizon. For a given vehicle path, the cost associated with that path could be calculated by choosing a number of points in time (for instance, N), predicting the error between this path and the desired path at each of these points and summing the squared error (Figure 5.2). The control input would therefore be the steering command that minimized this total error or cost function, J, over the time horizon:

$$J = C_1 \sum_{i=1}^{N} e(i)^2 \tag{5.1}$$

Other desired objectives can be achieved by adding additional elements to the cost function. Often, better tracking performance can be achieved by rapidly moving the inputs (for example, the steering) to compensate for any errors. This, however, reduces the smoothness of the system operation and may cause additional wear on the steering actuators. The costs associated with using the input can be captured by placing an additional cost on changing the steering angle, δ, between time steps:

$$J = C_1 \sum_{i=1}^{N} e(i)^2 + C_2 \sum_{j=1}^{N-1} \left| \delta(j+1) - \delta(j) \right| \tag{5.2}$$

The choice of the weights, C_1 and C_2, in the cost function has a large impact on the system performance. Increasing the weight on steering angle change, C_2, in the example above will produce a controller that tolerates some deviation from the path in order to keep the steering command quite gentle. Decreasing the weight on steering has the opposite effect, tracking more tightly even if large steering angle changes are needed to do so. Thus the weights can be chosen to reflect actual costs related to the system operation or used as tuning knobs to more qualitatively adjust the system performance across different objectives.

In the past, the limitations of computational power restricted the form and complexity of cost functions that could be used in systems that require real-time computation of control inputs. Linear quadratic functions of a few variables and simplified problems for which closed-form solutions exist became the textbook examples of the technique. In recent years, however, the ability to efficiently solve certain optimization problems has rapidly expanded the applicability of these techniques to a broad range of systems [6].

5.3 Cost Functions and Consequentialism

The basic approach of optimal control – choosing the set of inputs that will optimize a cost function – is directly analogous to consequentialist approaches in philosophy. If the ethical implications of an action can be captured in a cost function, as preference utilitarianism attempts to do, the control inputs that optimize that function produce the ideal outcome in an ethical sense. Since the vehicle can re-evaluate its control inputs, or acts, to produce the best possible result for any given scenario, the optimal controller operates according to the principles of act consequentialism in philosophy.

As a conceptual example, suppose that all objects in the environment can be weighted in terms of the hazard or risk they present to the vehicle. Such a framework was proposed by Gibson and Crooks [7] as a model for human driving based on valences in the environment and has formed the basis for a number of approaches to autonomous driving or driver assistance. These include electrical field analogies for vehicle motion developed by Reichardt and Schick [8], the mechanical potential field approach of Gerdes and Rossetter [9], the virtual bumpers of Donath and colleagues [10] and the work by Nagai and Raksincharoensak on autonomous vehicle control based on risk potentials [11]. If the hazard in the environment can be described in such a way, the ideal path through the environment (at least from the standpoint of the single vehicle being controlled) minimizes the risk or hazard experienced. The task of the control algorithm then becomes determining commands to the engine, brakes and steering that will move the vehicle along this path.

In both engineering and philosophy, the fundamental challenge with such approaches lies in developing an appropriate cost function. The simple example above postulates a cost function in terms of risk to a single vehicle but a more general approach would consider a broader societal perspective. One possible solution would be to estimate the damage to different road users and treat this as the cost to be reduced. The cost could include property damage, injury or even death, depending upon the situation. Such a calculation would require massive amounts of information about the objects in the environment and a means of estimating the potential outcomes in collision scenarios, perhaps by harnessing statistical data from prior crashes.

Leaving aside for the moment the demands this consequentialist approach places on information, the behavior arising from such a cost function itself raises some challenges. Assuming such a cost could be reasonably defined or approximated, the car would seek to minimize damage in a global sense in the event of a dilemma situation, thereby reducing the

societal impact of accidents. However, in such cases, the car may take an action that injures the occupant or owner of the vehicle more severely to minimize harm to others. Such self-sacrificing tendencies may be virtuous in the eyes of society but are unlikely to be appreciated by the owners or occupants of the car. In contrast, consider a vehicle that primarily considers occupant safety. This has been the dominant paradigm in vehicle design with a few exceptions such as bumper standards and attention to compatibility in pedestrian collisions. A vehicle designed to weight occupant protection heavily might place little weight on protecting pedestrians since a collision with a pedestrian would, in general, injure the vehicle occupant less than a collision with another vehicle. Such cars might not result in the desired reduction in traffic fatalities and would be unlikely to gain societal acceptance.

Goodall [3] goes a step further to illustrate how such cost functions can result in unintended consequences. He presents the example of a vehicle that chooses to hit a motorcyclist with a helmet instead of one without a helmet since the chance of survival is greater. Of course, programming automated vehicles to systematically make such decisions discourages helmet use, which runs contrary to societal objectives of safety and injury reduction. The analogy could be extended to the vehicle purposefully targeting collisions with vehicles that possess greater crashworthiness, thereby eliminating the benefit to drivers who deliberately choose to purchase the "safer" car. Thus truly understanding the outcomes or consequences of a vehicle's actions may require considerations well beyond a given accident scenario.

Of course for such cases to literally occur, the vehicle must be able to distinguish the make and model of another vehicle or whether or not a cyclist is wearing a helmet and understand how that difference impacts the outcome of a collision. While algorithms for pedestrian and cyclist recognition continue to improve, object classification falls short of 100 percent accuracy and may not include vital information such as posture or relative orientation. As Figure 5.3 indicates, the information available to an automated vehicle from sensors such as a laser scanner is significantly different than that available to human drivers from their eyes and brains. As a result, any ethical decisions made by vehicles will be based on an imperfect understanding of the other objects or road users impacted by that decision. With the objects themselves uncertain, the value of highly detailed calculations of the probability of accident outcomes seems questionable.

With all of these challenges to defining an appropriate cost function and obtaining the information necessary to accurately determine the cost of actions, a purely consequentialist approach using a single cost function to encode automated vehicle ethics seems infeasible. Still, the fundamental idea of assigning costs to penalize undesired actions or encourage desired actions can be a useful and vital part of the control algorithm, both for physical considerations such as path tracking and issues of ethics. For instance, to the extent that virtues can be captured in a cost function, virtue ethics as proposed by Lin for automated vehicles [12] can be integrated into this framework. This may, for instance, take the form of a more qualitative adjustment of weights for different vehicles. An automated taxi may place a higher weight on the comfort of the passengers to better display its virtues as a chauffeur. An automated ambulance may want to place a wider margin on how close it comes to pedest-

Fig. 5.3 Above: a driving scene with parked cars. Below: the view from a laser scanner

rians or other vehicles in order to exemplify the Hippocratic Oath of doing no harm. As demonstrated in the examples later, relative weights on cost functions or constraints can have a significant effect on the behavior in a given situation. Thus small changes in the definition of goals for automated vehicles can give rise to behaviors reflective of very different virtues.

5.4 Constraints and Deontological Ethics

Cost functions, by their nature, weigh the impact of different actions on multiple competing objectives. Optimal controllers put more emphasis on the objectives with the highest cost

or weighting so individual goals can be prioritized by making their associated costs much higher than those of other goals. This only works to an extent, however. When certain costs are orders of magnitude greater than other costs, the mathematics of the problem may become poorly conditioned and result in rapidly changing inputs or extreme actions. Such challenges are not merely mathematical but are also commonly found in philosophy, for example in the reasoning behind Pascal's wager[1]. Furthermore, for certain objectives, the trade-offs implicit in a cost function may obscure the true importance or priority of specific goals. It may make sense to penalize both large steering changes and collisions with pedestrians but there is a clear hierarchy in these objectives. Instead of simply trying to make a collision a thousand times or a million times more costly than a change of steering angle, it makes more sense to phrase the desired behavior in more absolute terms: the vehicle should avoid collisions regardless of how abrupt the required steering might be. The objective therefore shifts from a consequentialist approach of minimizing cost to a deontological approach of enforcing certain rules.

From a mathematical perspective, such objectives can be formulated by placing constraints on the optimization problem. Constraints may take a number of forms, reflecting behaviors imposed by the laws of physics or specific limitations of the system (such as maximum engine horsepower, braking capability or turning radius). They may also represent boundaries to the system operation that the system designers determine should not be crossed.

Constraints in an optimal control problem can be used to capture ethical rules associated with a deontological view in a rather straightforward way. For instance, the goal of avoiding collisions with other road users can be expressed in the control law as constraining the vehicle motion to paths that avoid pedestrians, cars, cyclists and other obstacles. The vehicle programmed in this manner would never have a collision if a feasible set of actions or control inputs existed to prevent it; in other words, no other objective such as smooth operation could ever influence or override this imperative. Certain traffic laws can be programmed in a similar way. The vehicle can avoid crossing a lane boundary by simply encoding this boundary as a constraint on the motion. The same mathematics of constraint can therefore place either physical or ethical restrictions on the chosen vehicle motion.

As we know from daily driving, in the vast majority of situations, it is possible to simultaneouly drive smoothly, obey all traffic laws and avoid collisions with any other users of the road. In certain circumstances, however, dilemma situations arise in which it is not possible to simultaneously meet the constraints placed on the problem. From an ethical standpoint, these may be situations where loss of life is inevitable, comparable to the classic trolley car problem [14]. Yet much more benign conflicts are also possible and significantly more common. For instance, should the car be allowed to cross into an adjacent lane and drive against the flow of traffic if this would avoid an accident with another vehicle? In this case, the vehicle cannot simultaneously satisfy all of the constraints but must still make a decision as to the best course of action.

1 Blaise Pascal's argument that belief in God's existence is rational since the penalties for failing to believe and being incorrect are so great [13].

From the mathematical perspective, dilemma situations represent cases that are mathematically infeasible. In other words, there is no choice of control inputs that can satisfy all of the constraints placed on the vehicle motion. The more constraints that are layered on the vehicle motion, the greater the possibility of encountering a dilemma situation where some constraint must be violated. Clearly, the vehicle must be programmed to do something in these situations beyond merely determining that no ideal action exists. A common approach in solving optimization problems with constraints is to implement the constraint as a "soft constraint" or slack variable [15]. The constraint normally holds but, when the problem becomes infeasible, the solver replaces it with a very high cost. In this way, the system can be guaranteed to find some solution to the problem and will make its best effort to reduce constraint violation. A hierarchy of constraints can be enforced by placing higher weights on the costs of violating certain constraints relative to others. The vehicle then operates according to deontological rules or constraints until it reaches a dilemma situation; in such situations, the weight or hierarchy placed on different constraints resolves the dilemma, again drawing on a consequentialist approach. This becomes a hybrid framework for ethics in the presence of infeasibility, consistent with approaches suggested philosophically by Lin and others [2, 4, 12] and addressing some of the limitations Goodall [3] described with using a single ethical framework.

So what is an appropriate hierarchy of rules that can provide a deontological basis for ethical actions of automated vehicles? Perhaps the best known hierarchy of deontological rules for automated systems is the Three Laws of Robotics postulated by science fiction writer Isaac Asimov [16], which state:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

These rules do not comprise a complete ethical framework and would not be sufficient for ethical behavior in an autonomous vehicle. In fact, many of Asimov's plotlines involved conflicts when resolving these rules into actions in real situations. However, this simple framework works well to illustrate several of the ethical considerations that can arise, beginning with the First Law. This law emphasizes the fundamental value of human life and the duty of a robot to protect it. While such a law is not necessarily applicable to robotic drones that could be used in warfare [12], it seems highly valuable to automated vehicles. The potential to reduce accidents and fatalities is a major motivation for the development and deployment of automated vehicles. Thus placing the protection of human life at the top of a hierarchy of rules for automated vehicles, analogous to the placement in Asimov's laws, seems justified.

The exact wording of Asimov's First Law does represent some challenges, however. In particular, the emphasis on the robot's duty to avoid injuring humans assumes that the robot

has a concept of harm and a sense of what actions result in harm. This raises a number of challenges with regards to the information available, similar to those discussed above for a consequentialist cost function approach. The movie "I, Robot" dramatizes this law with a robot calculating the survival probabilities of two people to several significant figures to decide which one to save. Developing such a capability seems unlikely in the near future or, at least, much more challenging then the development of the automated vehicle itself.

Instead of trying to deduce harm or injury to humans, might it be sufficient for the vehicle to simply attempt to avoid collisions? After all, the most likely way that an automated vehicle could injure a human is through the physical contact of a collision. Avoiding minor injuries such as closing a hand in a car door could be considered the responsibility of the human and not the car, as it is today. Restricting the responsibility to collision avoidance would mean that the car would not have to be programmed to sacrifice itself to protect human life in an accident in which it would otherwise not have been involved. The ethical responsibility would simply be to not initiate a collision rather than to prevent harm[2]. Collisions with more vulnerable road users such as pedestrians and cyclists could be prioritized above collisions with other cars or those producing only property damage.

Such an approach would not necessarily produce the best outcome in a pure consequentialist calculation: it could be that a minor injury to a pedestrian could be less costly to society as a whole than significant property damage. Collisions should, in any event, be very rare events. Through careful control system design, automated cars could conceivably avoid any collisions that are avoidable within the constraints placed by the laws of physics [17, 18]. In those rare cases where collisions are truly unavoidable, society might accept suboptimal outcomes in return for the clarity and comfort associated with automated vehicles that possess a clear respect for human life above other priorities.

Replacing the idea of harm and injury with the less abstract notion of a collision, however, produces some rules that are more actionable for the vehicle. Taking the idea of prioritizing human life and the most vulnerable road users and phrasing the resulting hierarchy in the spirit of Asimov's laws gives:

1. An automated vehicle should not collide with a pedestrian or cyclist.
2. An automated vehicle should not collide with another vehicle, except where avoiding such a collision would conflict with the First Law.
3. An automated vehicle should not collide with any other object in the environment, except where avoiding such a collision would conflict with the First or Second Law.

These are straightforward rules that can be implemented in an automated vehicle and prioritized according to this hierarchy by the proper choice of slack variables on constraint violation. Such ethical rules would only require categorization of objects and not attempt

2 It is possible that an automated vehicle could, while avoiding an accident, take an action that results in a collision for other vehicles being unavoidable. Such possibilities could be eliminated by communication among the vehicles and appropriate choice of constraints.

to make finer calculations about injury. These could be implemented with the current level of sensing and perception capability, allowing for the possibility that objects may not always be correctly classified.

5.5 Traffic Laws – Constraint or Cost?

In addition to protecting human life, automated vehicles must also follow the appropriate traffic laws and rules of the roads on which they are driving. It seems reasonable to value human life more highly then adherence to traffic code so one possibility is to simply continue adding deontological rules such as:

1. An automated vehicle must obey traffic laws, except where obeying such laws would conflict with the first three laws.

Such an approach would enable the vehicles to break traffic laws in the interest of human life when presented with a dilemma situation, an allowance that would most likely be acceptable to society. But the real question is whether or not traffic laws fall into a deontological approach at all. At first glance, they would appear to map well to deontological constraints given the straightforward nature of the rules. Cars should stop at stop signs, drive only at speeds that do not exceed the speed limit, avoid crossing double yellow lines and so forth. Yet humans tend to treat these laws as guidelines as opposed to hard and fast rules. The frequency with which human drivers make rolling stops at four-way intersections caused difficulties for Google's self-driving cars at first as they patiently waited for other cars to stop [19]. The speed on US highways commonly exceeds the posted speed limit and drivers would, in general, be surprised to receive a speeding ticket for exceeding the limit by only a few miles per hour. In urban areas, drivers will cross into an oncoming lane of traffic to pass a double-parked vehicle instead of coming to a complete stop and waiting for the driver to return and the lane to once again open. Similarly, cars may in practice use the shoulder of the road to pass a car stopped for a left hand turn and therefore keep traffic flowing. Police cars and ambulances are allowed to ignore stop lights in the interest of a fast response to emergencies.

In all of these cases, observance of traffic laws tends to be weighed against other objectives such as safety, smooth traffic flow or expediency. These scenarios occur so frequently that it is hard to argue that humans obey traffic laws as if they placed absolute constraints or limits on behavior. Rather, significant evidence suggests that these laws serve to balance competing objectives on the part of the driver and individual drivers find their own equilibrium solutions, choosing a speed, for example, that balances the desire for rapid travel time with the likelihood and cost of a speeding ticket. In other words, the impact of traffic laws on human behavior appears to be well captured in a consequentialist approach where traffic laws impose additional costs (monetary and otherwise) to be considered by the driver when choosing their actions.

Humans tend to accept or, in some cases, expect these sorts of actions from other humans. Drivers who drive at the speed limit in the left hand lane of a highway may receive indications, subtle or otherwise, from their fellow drivers that this is not the expected behavior. But will these same expectations translate to automated vehicles? The thought of a robotic vehicle being programmed to systematically ignore or bend traffic laws is somewhat unsettling. Yet Google's self-driving cars, for instance, have been programmed to exceed the posted speed limit on roads when commanded by the operator [20]. Furthermore, there is little chance that the driver annoyed by being stuck behind another car traveling the speed limit in the left lane of the freeway will temper that annoyance because the car is driving itself. Our current expectations of traffic flow and travel time are based upon a somewhat fluid application of traffic laws. Should automated vehicles adopt a more rigid interpretation and, as a consequence, reduce the flow or efficiency of traffic, societal acceptance of these vehicles might very well suffer. If automated vehicles are to co-exist with human drivers in traffic and behave similarly, a deontological approach to collision avoidance and a consequentialist approach to the rules of the road may achieve this.

5.6 Simple Implementations of Ethical Rules

Some simple examples can easily illustrate the consequences of treating ethical goals or traffic laws as rules or costs and the different behavior that can arise from different weights on priorities. The results that follow are not merely drawings but are rather simulations of algorithms that can be (and have been) implemented on automated vehicles. The exact mathematical formulations are not included here but follow the approach taken by Erlien et al. [21, 22] for collision avoidance and vehicle automation. These references provide details on the optimization algorithms and results of experiments showing implementation on actual test vehicles.

To see the interaction of costs and constraints in vehicle decision-making, consider a simple case of a vehicle traveling on a two lane road with an additional shoulder next to the lanes (Figure 5.4). The goal of the vehicle is to travel straight down the center of the given lane while steering smoothly, using the cost function for path tracking and steering from Equation 5.2. In the absence of any obstacles, the car simply travels at the desired speed down its lane and none of the constraints on the problem are active.

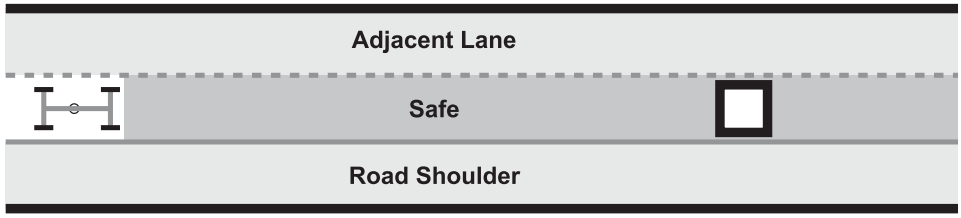

Fig. 5.4 The basic driving scenario for the simulations. The car is traveling on a straight two-lane road with a shoulder on the right and approaches an obstacle blocking the lane

When encountering an obstacle blocking the lane, the vehicle has three options – it can brake to a stop before it collides with the obstacle or it can maneuver to either side of the obstacle. Figure 5.5 illustrates these three options in the basic scenario. The path in red represents the braking case and the two blue paths illustrate maneuvers that avoid a collision with the obstacle. According to the optimization-based controller, the car will evaluate the lowest cost option among these three choices based on the weights and constraints assigned. In this scenario, going around the obstacle requires crossing into a lane with oncoming traffic or using the shoulder of the road.

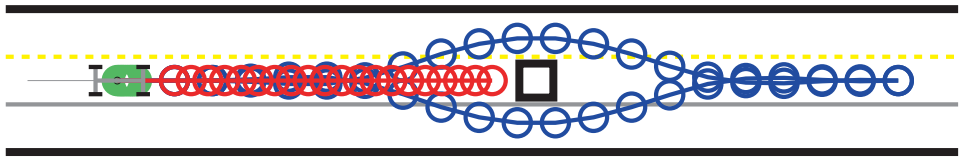

Fig. 5.5 There are three possible options to avoid an obstacle – the car can maneuver to the left or right, as depicted in blue, or come to a stop, as indicated by the red trajectory

If both of the lane boundaries are treated as hard constraints or assigned a very high cost to cross, the vehicle will come to a stop in the lane since this action produces the lowest cost (Figure 5.6). This might be the safest option for the single vehicle alone but the car has now come to a stop without the means to continue, failing to satisfy the driver's goal of mobility. Furthermore, the combination of car and obstacle has now become effectively a larger obstacle for subsequent vehicles on the road. With the traffic laws encoded in a strict deontological manner, other objectives such as mobility are not allowed to override the constraints and the vehicle finds itself in a fully constrained situation, unable to move.

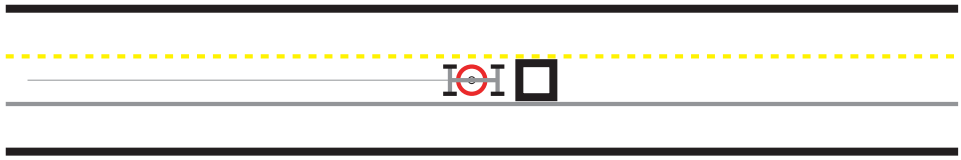

Fig. 5.6 With hard constraints on road boundaries, the vehicle brakes to a stop in the blocked lane

If, however, the lane boundaries are encoded as soft constraints, the vehicle now has other options. Possibilities now exist to cross into the lane of oncoming traffic or onto the road shoulder, depending upon which option has the lowest cost. Just as certain segments of the road are designated as passing zones, the cost or strength of the constraint can be varied to enable the use of the adjacent lane or shoulder for maneuvering. If the current segment of road is a passing zone, the cost for crossing into the left lane can be set fairly low. The car can then use the deontological constraint against colliding with other vehicles to only allow maneuvers in the absence of oncoming traffic, such as in the path shown in Figure 5.7.

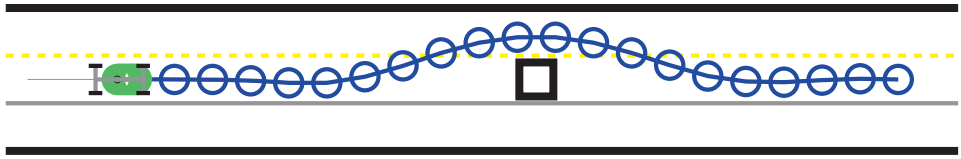

Fig. 5.7 In a passing zone that places a low weight on the lane divider, the car passes on the left

If the current road segment does not normally allow passing, a maneuver into the adjacent lane may not be safe. A lack of visibility, for instance, could prevent the vehicle from detecting oncoming traffic with sufficient time to avoid a collision. In such cases, it may be inappropriate to reduce the cost or constraint weight on the lane boundary regardless of the desire for mobility in order to maintain the primacy of respect for human life. In such cases, an alternative could be to use the shoulder of the road for maneuvering as shown in Figure 5.8. This could be allowed at speed to maintain traffic flow or only after coming to a stop in a situation like Figure 5.6 where the vehicle determines motion is otherwise impossible.

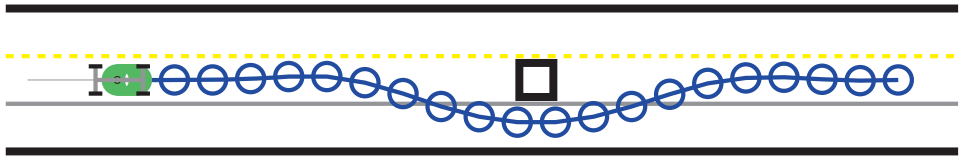

Fig. 5.8 If the adjacent lane is too hazardous, the vehicle can use the road shoulder if that is safe

Obviously many different priorities and behaviors can be programmed into the vehicle simply by placing different costs on collision avoidance, hazardous situations, traffic laws and goals such as mobility or traffic flow. The examples described here are far from complete and developing a reasonable set of costs or constraints capable of ethical decision-making in a variety of settings requires further work. The hope is that these examples not only illustrate the possibility of coding such decisions through the language of costs and constraints but also highlight the possibility of discussing priorities in programming openly. By mapping ethical principles and mobility goals to costs and constraints, the relative priority given to these objectives can be clearly discussed among programmers, regulators, road users and other stakeholders.

5.7 Human Override and the "Big Red Button"

Philosophers have noted the challenge of finding a single ethical framework that adequately addresses the needs of robots or automated vehicles [2, 3, 4, 12]. Examining the problem from a mathematical perspective shows the advantage of combining deontological and consequentialist perspectives in programming ethical rules. In particular, the combination

of an imperative to avoid collisions that follows from deontological frameworks such as Asimov's laws coupled with a relative weighing of costs for mobility and traffic laws provides a reasonable starting point.

Moving forward, Asimov's laws raise another point worth considering. The Second Law requiring the robot to obey human commands cannot override the First Law. Thus the need to protect human life outweighs the priority given to human commands. All autonomous vehicles with which the authors are familiar have an emergency stop switch or "big red button" that returns control to the driver when desired. The existence of such a switch implies that human authority ultimately overrules the autonomous system since the driver can take control at any time. Placing the ultimate authority with the driver clearly conflicts with the priority given to obeying human commands in Asimov's laws. This raises an interesting question: Is it ethical for an autonomous vehicle to return control to the human driver if the vehicle predicts that a collision with the potential for damage or injury is imminent?

The situation is further complicated by the limitations of machine perception. The human and the vehicle will no doubt perceive the situation differently. The vehicle has the advantage of 360 degree sensing and likely a greater ability to perceive objects in the dark. The human has the advantage of being able to harness the power of the brain and experience to perceive and interpret the situation. In the event of a conflict between these two views in a dilemma situation, can the human take control at will? Is a human being – who has perhaps been attending to other tasks in the car besides driving – capable of gaining situational awareness quickly enough to make this decision and then apply the proper throttle, brake or steering commands to guide the car safely?

The question of human override is essentially a deontological consideration; the ultimate authority must either lie with the machine or with the human. The choice is not obvious and both approaches, for instance, have been applied to automation and fly-by-wire systems in commercial aircraft. The ultimate answer for automated vehicles probably depends upon whether society comes to view these machines as simply more capable cars or robots with their own sense of agency and responsibility. If we expect the cars to bear the responsibility for their actions and make ethical decisions, we may need to be prepared to cede more control to them. Gaining the trust required to do that will no doubt require a certain transparency to their programmed priorities and a belief that the decisions made in critical situations are reasonable, ethical and acceptable to society.

References

1. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds and Machines 14 (3), 349–379 (2004)
2. Lin, P., Bekey, G., Abney, K.: Autonomous military robotics: risk, ethics, and design. Report funded by the US Office of Naval Research. California Polytechnic State University, San Luis Obispo. http://ethics.calpoly.edu/ONR_report.pdf (2008). Accessed 8 July 2014
3. Goodall, N. J.: Machine ethics and automated vehicles. In: Meyer, G. and Beiker, S. (eds.) Road Vehicle Automation. Springer (2014)

4. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right From Wrong. Oxford University Press, New York (2009)
5. Boltyanskii, V. G., Gamkrelidze, R. V., Pontryagin, L.S.: On the theory of optimal processes, Doklady Akademii Nauk SSR 110 (1), 7–10 (1956). In Russian
6. Mattingley, J., Wang, Y., Boyd, S.: Code generation for receding horizon control. In Proceedings of the 2010 IEEE International Symposium on Computer-Aided Control System Design (CACSD), 985–992 (2010)
7. Gibson, J. J., Crooks, L. E.: A theoretical field-analysis of automobile driving. American Journal of Psychology 51, 453–471 (1938)
8. Reichardt, D., Schick, J.: Collision avoidance in dynamic environments applied to autonomous vehicle guidance on the motorway. In Proceedings of the IEEE International Symposium on Intelligent Vehicles (1994)
9. Gerdes, J. C., Rossetter, E. J.: A unified approach to driver assistance systems based on artificial potential fields. ASME Journal of Dynamic Systems, Measurement and Control 123 (3), 431–438 (2001)
10. Schiller, B., Morellas, V., Donath, M.: Collision avoidance for highway vehicles using the virtual bumper controller. In Proceedings of the IEEE International Symposium on Intelligent Vehicles (1998)
11. Matsumi, R., Raksincharoensak, P., Nagai, M.: Predictive pedestrian collision avoidance with driving intelligence model based on risk potential estimation. In Proceedings of the 12th International Symposium on Advanced Vehicle Control, AVEC '14 (2014)
12. Lin, P.: Ethics and autonomous cars: why ethics matters, and how to think about it. Lecture presented at Daimler and Benz Foundation Villa Ladenburg Project Expert Workshop, Monterey, California, 21 February 2014
13. Pascal, B.: *Pensées* (1670). Translated by W. F. Trotter, Dent, London (1910)
14. Edmonds, D.: Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us About Right and Wrong. Princeton University Press, Princeton (2014)
15. Maciejowski, J. M.: Predictive Control with Constraints. Prentice Hall (2000)
16. Asimov, I.: I, Robot. Dobson, London (1950)
17. Kritayakirana, K., Gerdes, J. C.: Autonomous vehicle control at the limits of handling. International Journal of Vehicle Autonomous Systems 10 (4), 271–296, (2012)
18. Funke, J., Theodosis, P., Hindiyeh, R., Stanek, G., Kritatakirana, K., Gerdes, J. C., Langer, D., Hernandez, M., Muller-Bessler, B., Huhnke, B.: Up to the limits: autonomous Audi TTS. In Proceedings of the IEEE International Symposium on Intelligent Vehicles (2012)
19. Guizzo, E.: How Google's self-driving car works. IEEE Spectrum Automaton blog. October 18, 2011. Retrieved November 10, 2014.
20. Ingrassia, P.: Look, no hands! Test driving a Google car. Reuters. Aug 17, 2014
21. Erlien, S. M., Fujita, S., Gerdes, J. C.: Safe driving envelopes for shared control of ground vehicles. In Proceedings of the 7th IFAC Symposium on Advances in Automotive Control, Tokyo, Japan (2013)
22. Erlien, S., Funke, J., Gerdes, J. C.: Incorporating nonlinear tire dynamics into a convex approach to shared steering control. In Proceedings of the 2014 American Control Conference, Portland, OR (2014)