Ethical Decision Making During Automated Vehicle Crashes

Noah J. Goodall

Automated vehicles have received much attention recently, particularly the Defense Advanced Research Projects Agency Urban Challenge vehicles, Google's self-driving cars, and various others from auto manufacturers. These vehicles have the potential to reduce crashes and improve roadway efficiency significantly by automating the responsibilities of the driver. Still, automated vehicles are expected to crash occasionally, even when all sensors, vehicle control components, and algorithms function perfectly. If a human driver is unable to take control in time, a computer will be responsible for precrash behavior. Unlike other automated vehicles, such as aircraft, in which every collision is catastrophic, and unlike guided track systems, which can avoid collisions only in one dimension, automated roadway vehicles can predict various crash trajectory alternatives and select a path with the lowest damage or likelihood of collision. In some situations, the preferred path may be ambiguous. The study reported here investigated automated vehicle crashing and concluded the following: (*a*) automated vehicles would almost certainly crash, (*b*) an automated vehicle's decisions that preceded certain crashes had a moral component, and (*c*) there was no obvious way to encode complex human morals effectively in software. The paper presents a three-phase approach to develop ethical crashing algorithms; the approach consists of a rational approach, an artificial intelligence approach, and a natural language requirement. The phases are theoretical and should be implemented as the technology becomes available.

Automated vehicle technologies are the computer systems that assist human drivers by automating aspects of vehicle control. These technologies have a range of capabilities, from antilock brakes and forward collision warning, to adaptive cruise control and lane keeping, to fully automated driving. NHTSA has defined five levels of vehicle automation, described in Table 1 (*1*). The study reported in this paper focused on automation Levels 3 and 4, with some applications in Level 2. Throughout this paper, the term "automated vehicles" refers to vehicles with these high-level capabilities.

Automated vehicles have been operating on public roadways since 1994, with the field-testing of Daimler–Benz's VITA-2 and Universität der Bundeswehr München's VaMP in Europe (*2*). The following year, Carnegie Mellon's Navlab 5 vehicle (*3*) demonstrated autonomous steering in the United States (*4*), while the University of Parma's automated vehicle ARGO completed autonomous trips in Italy in 1998 (*5*). The U.S. Department of Transportation formed the National Automated Highway System Consortium to prototype fully automated highway driving, which led to a 1997 demonstration of vehicle platooning and automated driving aided by embedded magnets and radar-reflective tape in the roadway (*6*). Automated vehicles from three teams completed the Defense Advanced Research Projects Agency Grand Challenge of 2007 by navigating a complex urban environment within a time limit (*7*). In 2010, Google announced that it had been testing a fleet of seven automated vehicles on public roadways, with more than 140,000 mi driven with occasional human intervention (*8*). Several major automakers have since announced research efforts, including Audi (*9*), Ford (*10*), BMW (*11*), Mercedes-Benz (*12*), General Motors (*13*), and Toyota (*14*).

Virginia Center for Transportation Innovation and Research, 530 Edgemont Road, Charlottesville, VA 22903. noah.goodall@vdot.virginia.gov.

Transportation Research Record: Journal of the Transportation Research Board, No. 2424, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 58–65.
DOI: 10.3141/2424-07

OBJECTIVE

Although a great deal of work has been done in road vehicle automation and obstacle avoidance, no known research has been published on optimal crashing strategies for Levels 3 and 4 automated vehicles. Of the existing laws that govern automated vehicle behavior within the United States, none addresses computerized control of precrash or crash-avoidance behavior. Instead they require a human to be available to take control of the vehicle without notice (e.g., Vehicles with Autonomous Technology, Florida CS/HB 1207, 2012; Autonomous Vehicles: Safety Requirements, Florida SB 1298, 2012; Autonomous Vehicle Act of 2012, Council of the District of Columbia, B19-0913, 2012). The objective of this study was to assess the need for a moral component to automated vehicle decision making during unavoidable crashes, and to identify the most promising strategies from the field of machine ethics for application in road vehicle automation.

The remainder of this paper focuses on three arguments: that even perfectly functioning automated vehicles will crash, that certain crashes require the vehicle to make complex ethical decisions, and that there is no obvious way to encode human ethics in computers. Finally, an incremental, hybrid approach to develop ethical automated vehicles is proposed.

AUTOMATED VEHICLE CRASH POTENTIAL

Much of the excitement that surrounds automated vehicles seems to have its basis in the assumption that they will be safer than vehicles driven by human drivers. The empirical evidence does not refute this claim: Google self-driving cars had traveled more than 435,000 mi on public roads as of April 2013 (*15*) with only one crash (*16*), which Google claims occurred while the vehicle was under the control of a human driver (*17*). This mileage was not accrued by unassisted automated vehicles alone. The self-driving cars were tightly supervised by test drivers, who were required to

TABLE 1 NHTSA Road Vehicle Automation Levels (*1*)

NHTSA Automation Level	Description
0. No automation	Driver is in complete control of steering, braking, and throttle, although vehicle may provide warnings.
1. Function-specific automation	Vehicle may independently automate one or more control functions.
2. Combined-function automation	At least two control functions are automated and operated in conjunction (e.g., adaptive cruise control and lane centering). Driver may have to take control with no notice.
3. Limited self-driving automation	Driver can cede full control to the vehicle in some situations, and driver has a reasonable amount of transition time before he or she must take control.
4. Full self-driving automation	Vehicle can safely pilot the vehicle for an entire trip, with no expectation for the driver to take control. Such a vehicle does not yet exist.

intervene to prevent hazardous situations. Those close to the project have stated that the Google vehicles can travel 50,000 miles on freeways without intervention (*18*).

Automated vehicles cannot yet claim to be significantly safer than vehicles driven by humans. Smith noted that, with a Poisson distribution and national mileage and crash estimates, an automated vehicle would need to drive 725,000 mi on representative roadways without incident and without human assistance to say with 99% confidence that they crashed less frequently than vehicles with human drivers, and they would have to drive 300 million mi if only fatal crashes were considered (*19*). An automated heavy truck would need to travel 2.6 million mi without crashing to demonstrate safety benefits, compared with a truck driven by a human at 99% confidence, and 241 million mi without a fatality, given that today's trucks employ professional, trained drivers and a great portion of their mileage is driven on relatively safe freeways. An automated vehicle has yet to travel these distances unassisted. Relevant calculations are summarized in Table 2.

Crashes may occur as a result of software or hardware failures (e.g., loss of steering control on a curve). An automated vehicle sold to the public would likely require multiple redundancies, extensive testing, and frequent mandatory maintenance to minimize these types of failures. Any engineering system can fail. As for an automated vehicle, even a perfectly functioning system cannot avoid every collision.

Theoretical research robotics confirms that a crash-free environment is unrealistic. Although many techniques are described in the literature on collision avoidance in a dynamic environment (*23*), most are unable to ensure the avoidance of collisions with unpredictable objects. In response, Fraichard and Asama introduced the concept of an inevitable collision state for mobile robots, defined as "a state for which, no matter what the future trajectory of the system is, a collision with an obstacle eventually occurs" (*24*). To ensure a mobile robot's safety, its future trajectories are checked to ensure that at no point it enters an inevitable collision state (e.g., moves toward an obstacle without adequate time to brake). In the application of this concept to public roadways, the researchers acknowledged that, even with perfect sensing, it would be impossible to guarantee safety in the presence of unpredictable obstacles such as human-driven vehicles, pedestrians, and wildlife (*25*). Instead, a probability model of expected vehicle behavior was used to ensure that automated vehicles could avoid moderate- and high-risk situations (*26*). As of this writing, Google is attempting to patent this concept for automated vehicles (D. I. Ferguson and D. A. Dolgov, Modifying Behavior of Autonomous Vehicle Based on Predicted Behavior of Other Vehicles, U.S. Patent Application 20130261872 Kind Code: A1, Oct. 3, 2013). When human-driven vehicles deviate from these models and behave unexpectedly, crashes can occur.

It is not difficult to imagine a scenario in which a crash is unavoidable, even for an automated vehicle with complete knowledge of its world and negligible reaction time. For example, an automated vehicle is vulnerable when it is stopped at a traffic signal, surrounded on all sides by queued vehicles, while a distracted truck driver approaches from behind. It is impossible for the automated vehicle to avoid some type of collision, although it may use evasive maneuvers to minimize the impact.

Automated vehicle proponents often cite statistics that show a large percentage of vehicle crashes are at least partially the result of human error (*27*). This reality adds to the challenge of automated vehicle makers, because their vehicles will be forced to interact, for the foreseeable future, with occasionally dangerous human drivers, not to mention pedestrians, bicyclists, wildlife, and debris.

The superiority of automated vehicles to human drivers for safety has not been proven statistically. Even in simulations with perfect sensing, crashes are still possible. Given the dynamic environment of roadway traffic, proximity to vehicles with high-speed differentials, and limited maneuverability of automated vehicles at high speeds,

TABLE 2 Required Mileage of Automated Vehicles to Demonstrate Safety Benefits (*20–22*)

	All Vehicles		Heavy Trucks	
Calculated Item	All Crashes	Fatal Crashes	All Crashes	Fatal Crashes
Vehicle miles traveled (VMT)	$2{,}954 \times 10^9$	$2{,}954 \times 10^9$	168×10^9	168×10^9
Number of vehicles involved in crashes	18,705,600	45,435	295,900	3,200
VMT per crash	160,000	65,000,000	567,000	52,000,000
Crashless VMT required for benefit[a]	725,000	300,000,000	2,610,000	241,000,000

[a]Poisson distribution and p-value < .01 based on 2009 data.

automated vehicle safety cannot be ensured. Crash risk simply cannot be eliminated through more sophisticated algorithms or sensors.

DECISION MAKING DURING CRASHES

Shortcomings of Human Drivers

With today's vehicles, drivers in unsafe situations must make decisions to avoid collisions, and if a collision is unavoidable, to crash as safely as possible. These decisions are made quickly and under great stress, with little forethought or planning. Automated vehicles today rely on human drivers to take over if part of the system fails, or if the automated vehicle encounters a situation it does not understand, such as construction. A human is expected to be able to take control at any time and with no notice in Level 2 automated vehicles, and with reasonable notice in Level 3 vehicles (*1*).

Early research suggests that to expect humans to remain alert may be overly optimistic. In a recent study, participants that drove semi-automated vehicles on a closed test track engaged in eccentric head turns and secondary tasks, such as reading, significantly more than a control group, although effective countermeasures did reduce this behavior somewhat (*28*). If experienced drivers begin to rely on automation technologies, the next generation of drivers, which will have grown up around automated vehicles, will likely be even less vigilant in their monitoring of the roadway.

Adequate Warning

In Level 3 automation, the driver is not required to remain attentive but must be available to take control of the vehicle within a certain amount of time after the receipt of an alert. NHTSA's definition of Level 3 automation does not specify what length of time constitutes an adequate warning. AASHTO recommends a minimum distance to allow a driver to perceive an unusual situation and react, from between 200 and 400 m at speeds of 100 km/h, depending on the type of road and maneuver (*29*). This distance would have to be significantly longer if the object in question were approaching from the opposite direction, up to 800 m.

AASHTO's recommendations are meant for complex situations that often require lane changes (e.g., when a vehicle approaches a toll turnstile) and represent the upper end of required distances. Still, crashes can occur with little notice, and a human may not be able to assess the situation and make a better decision than a computer that has been continuously monitoring the roadway.

The warning system's sensitivity to perceived risk poses an additional challenge. In the early 2000s, the Virginia Tech Transportation Institute at Virginia Polytechnic Institute and State University equipped vehicles with many sensors and recording devices to study naturalistic driving behavior of test subjects on public roadways (*30*). In this study, thresholds were used to determine when a vehicle might have experienced a near-crash event. Certain safety maneuvers appeared identical to near crashes after the data from vehicle sensors were analyzed. For example, a maneuver known as a flying pass, in which a vehicle approaches a stopped queue at high speed and abruptly changes lanes into a dedicated left- or right-turn lane, often appeared indistinguishable from a near crash. Researchers were forced to analyze video footage of the driver's face for signs of alarm (*30*).

Automated vehicles that misunderstand a human driver's intentions in situations such as a flying pass can lead to false alarms because a safe situation is interpreted as a dangerous one. Over time, such instances could decrease a driver's vigilance. In a Level 3 automated vehicle, it may be unrealistic to ask a driver to take control of a vehicle less than a few seconds before a collision, even though this situation may occur often. For these reasons, it is likely that a computer will maintain control of an automated vehicle when it encounters dangerous situations and during crashes.

ETHICS OF CRASHING

Human drivers may often make poor decisions during and before crashes. They must overcome severe time constraints, limited experience with their vehicles at the limits of handling, and a narrow cone of vision. Although today's automated vehicles also have somewhat limited sensing and processing power, the focus of the study reported in this paper was on advanced vehicles with near-perfect systems. This focus anticipates the criticism that future sensors and algorithms will eliminate all crashes, and therefore obviate vehicle ethics research (*31*). If even perfect vehicles must occasionally crash, there will always be a need for some type of ethical decision-making system.

These advanced automated vehicles will be able to make precrash decisions with sophisticated software and sensors that can accurately detect nearby vehicle trajectories and perform high-speed avoidance maneuvers. Thus they will be able to overcome many of the limitations experienced by humans. If a crash is unavoidable, a computer can quickly calculate the best way to crash on the basis of a combination of safety, the likelihood of the outcome, and certainty in measurements much faster and with greater precision than a human can. The computer may decide that to brake alone is not optimal, because at highway speeds it is often more effective to brake and swerve, or even to swerve and accelerate in an evasive maneuver.

One major disadvantage of automated vehicles during crashes is that, unlike a human driver who can decide how to crash in real time, an automated vehicle's decision of how to crash was defined by a programmer ahead of time. The automated vehicle can interpret the sensor data and make a decision, but the decision itself is a result of logic developed and coded months or years ago. This process does not pose a problem in cases in which a crash can be avoided: the vehicle selects the safest path and proceeds. If, however, injury cannot be avoided, the automated vehicle must decide how best to crash. This decision quickly becomes a moral one, demonstrated in an example from Marcus (*32*), modified slightly for this study. An automated vehicle is traveling on a two-lane bridge when a bus that is traveling in the opposite direction suddenly veers into its lane (Figure 1). The automated vehicle must decide how to react with the use of whatever logic has been programmed in advance. The three alternatives are as follows:

1. Veer left and off the bridge, which guarantees a severe, one-vehicle crash;
2. Crash head-on into the bus, which will result in a moderate, two-vehicle crash; and
3. Attempt to squeeze pass the bus on the right. If the bus suddenly corrects back toward its own lane (a low-probability event given how far the bus has drifted) a crash is avoided. If the bus does not correct itself, a high-probability event, then a severe, two-vehicle crash results. This crash would be a small, offset crash, which carried a greater risk of injury than the full, frontal collision in Alternative 2 (*33*).

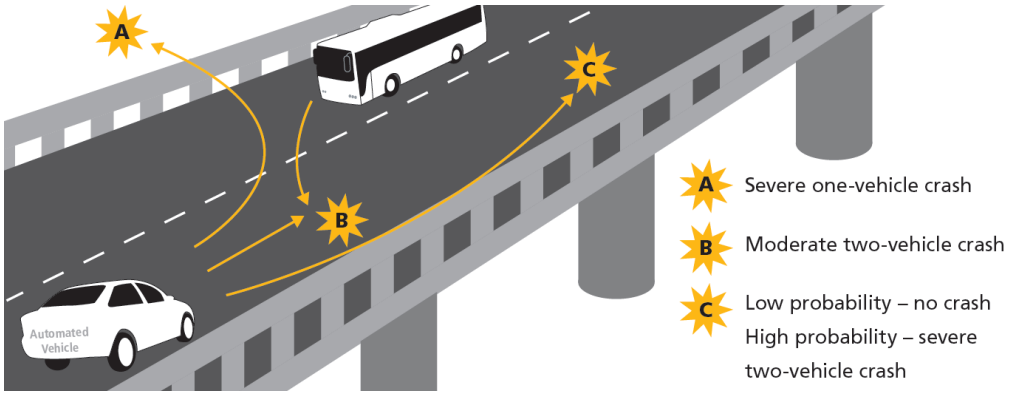

FIGURE 1 Diagram of three alternative trajectories for an automated vehicle when an oncoming bus suddenly enters the vehicle's lane.

These outcomes can be predicted only by the automated vehicle, and are not certain. The automated vehicle's path-planning algorithm would have to determine quickly the range of possible outcomes for each considered path, the likelihood of those outcomes occurring, and the algorithm's confidence in these estimates on the basis of the quality of sensor data and other factors. Finally, the algorithm would need to somehow optimize an objective function over the range of considered paths and quickly determine the safest route. A primitive form of this objective function could take the following form:

$$f(X) = \sum_{x \in X} s(x)\mathbf{P}(x|X)$$

where

X = set of possible outcomes x for a single trajectory choice,
s = function that represents the severity of a specific outcome x, and
$\mathbf{P}$ = conditional probability of an outcome occurring, given the trajectory choice.

The possible outcomes include property damage, injury, and death, each with a range of severities based on additional, unlisted factors. The calculation of the range of possible choices, outcomes, and probabilities of occurrence is the most technically complex part of the equation, whereas the calculation of the severity function s is the most morally difficult component. The reduction of death, injury, and property damage to a comparable number is problematic for many reasons, but without some type of ethical guidance, the automated vehicle has no real way to evaluate risk in the scenario shown in Figure 1.

ETHICAL VEHICLE DESIGN

There has been little discussion of the legal and moral implications of automated vehicle decision making during unavoidable crashes. Most of the research on moral machines has focused on military applications (*34, 35*) or general machine intelligence (*36*). A relatively recent area of study is machine ethics, which focuses on the development of autonomous machines that can exhibit moral behavior when they encounter new situations. In this section, an overview is presented of machine ethics and its applications to automated vehicles.

Rational Approaches

The instinct for engineers is to directly instruct the automated system how to behave in a variety of circumstances. This rationalist approach often takes the form of either deontology, in which the automated system must adhere to a set of rules, or consequentialism, in which the system's goal is to maximize some benefit. These rational approaches often appeal to engineers because computers can follow rules easily and maximize functions. Unfortunately, this strategy has several shortcomings, as described in the following examples.

Asimov's Laws of Robotics

When many first encounter machine ethics, they are reminded of Isaac Asimov's laws of robotics (*37*), reprinted here with references to computers replaced by automated vehicles.

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

Asimov developed these laws, not as a proposal for a rule-based robot code of ethics, but as a literary device to demonstrate the flaws in such a system. Machines are incredibly literal, and any system of ethics with only three rules requires that the follower possess some common sense or intuition, particularly if there is a conflict between the rules or within a single rule.

For example, the first rule forbids an automated vehicle to allow a human to "come to harm." A literal interpretation of this rule would prohibit sudden braking, even to avoid a collision, because it would cause whiplash to the vehicle occupants. Wallach and Allen provide an example of a life-saving surgery, during which a robot would be unable to "cause harm" and make the first incision (*38*). A driverless car that followed these rules might refuse to shift into gear: to leave the driveway and enter traffic would carry some risk of injury to a human.

Rules could be continuously added or clarified to cover these unique scenarios. Such a task would be somewhat futile. As vehicles continued to encounter novel situations, freak accidents, new road

designs, and vehicle system upgrades would need to be addressed. Someone would be required to decide with absolute moral authority what was right and wrong in every roadway situation. In the previous example shown in Figure 1, what if the bus had no passengers? What if it had one, or two, or three passengers? Would it matter if they were adults or children? Ethics as a field has not advanced to this point. As noted by Beavers (*39*), "after two millennia of moral inquiry, there is still no consensus on how to determine moral right and wrong." Although deontological ethics can provide guidance in many situations, it is not suitable as a complete ethical system because of the incompleteness of any set of rules and the difficulty involved in the articulation of complex human ethics as a set of rules.

Consequentialism

Asimov's laws may seem too abstract, given their science fiction background. A more familiar rational approach to automated vehicle ethics is consequentialism. In this system, a vehicle decides on a crash trajectory that minimizes global harm or damage. If the options are collision versus no collision, or property damage versus serious injury, the choice is obvious.

The way damage is calculated, however, can lead to undesirable outcomes. For example, how is harm or damage quantified? The most obvious way is in dollars, with damage estimates from the insurance industry. This method creates problems in practice. If the objective was to try and minimize cost, automated vehicles would choose to collide with the less expensive of the two vehicles, given a choice. If the collision was severe and injury likely, the automated vehicle would choose to collide with the vehicle with the higher safety rating, or choose to collide with a helmeted motorcyclist instead of a helmetless rider. Many would consider this decision unfair, not only because of discrimination but also because those who paid for safety were targeted while those who did not were spared. The fact that such decisions were made by a computer algorithm would be no consolation.

A second problem with consequentialism is in the determination of what information to incorporate into the decision and what to leave out. For example, crash compatibility measures the damage inflicted by one type of vehicle in collisions with another type (*40*). An automated vehicle could detect nearby vehicle types from its vision system and, in an unavoidable collision, attempt to collide with more compatible vehicles. Although this decision might improve safety, it would be morally unacceptable because safer vehicles were unfairly targeted.

Even if the two vehicles in a collision are identical, occupants experience different risks on the basis of demographics. Evans summarized an analysis of the Fatality Analysis Reporting System data set (*41*).

> If one driver is a man, and the other a similar-age woman, the woman is 28% more likely to die. If one driver is age 20 and the other age 70, the older driver is three times as likely to die. If one driver is drunk and the other sober, the drunk is twice as likely to die [because alcohol affects many body organs, not just the brain]. (*42*) If one driver is traveling alone while the other has a passenger, the lone driver is 14% more likely to die than the accompanied driver, because the accompanied driver is in a vehicle heavier by the mass of its passenger. (*43*)

An automated vehicle that used facial recognition technology could estimate the gender, age, or level of inebriation of nearby drivers and passengers, and adjust its objective function accordingly, although, again, many would probably find such estimates disturbing.

Both rational approaches to automated vehicle ethics discussed in this section demonstrate shortcomings. Deontological approaches require abstract rules that may not be applicable or may conflict in specific situations, while consequentialist approaches are so rigid that they produce actions that many consider reprehensible. These undesirable outcomes result from the inherent literalness of computers and from the inability of humans to articulate their own morals. For these reasons, rational approaches alone have limited applicability to automated vehicle ethics.

Artificial Intelligence Approaches

For years, automated language translation relied on rules developed by experts. The expectation was that language could be defined by rules, given enough time to learn the rules and write them out. An alternative approach with the use of algorithms that study and learn language automatically, without formal rules, has experienced much more success than rule-based methods (*44*). These techniques are known as artificial intelligence. Language translation offers a fitting analogy for ethical systems. In both areas, artificial intelligence methods are useful when the rules cannot be articulated.

Artificial intelligence methods have the potential to learn human ethics through the observation of human actions or through rewards for their own moral behavior. A computer can identify the components of ethics on its own, without the need for a human to articulate precisely why an action is or is not ethical. Wallach and Allen refer to these techniques as "bottom-up" approaches, which can include such techniques as genetic algorithms, connectionism, and learning algorithms (*38*). Artificial neural networks, which use layers of nodes in a connectionist computing approach to find complex relationships between inputs and outputs, have been used in a simple case to classify hypothetical decisions as either moral or amoral (*45*). Hibbard, in the formulation of a consequentialist approach to machine ethics, proposed a similar method by which an independent artificial intelligence agent calculated the moral weights assigned by humans after study participants were polled across a wide range of hypothetical situations (*46*). An early automated vehicle project, Carnegie Mellon's Autonomous Land Vehicle in a Neural Net, used a simple back-propagation-trained artificial neural network to teach itself to steer by watching a human driver for just 2 min (*47*). A similar technique could be used, with much more training data, to understand how humans choose to behave—or should behave—in morally complex driving situations when time is not a factor. The neural network could be trained on a combination of simulation and recordings of crashes and near crashes, with human feedback on the ethical response.

Artificial intelligence techniques have several shortcomings. If not carefully designed, they risk emulation of how humans behave rather than what they believe. For example, a human may choose to push a nearby vehicle into oncoming traffic to avoid his own collision. Self-preservation instincts that do not maximize overall safety may be realistic but not ethical. Ethics addresses how humans ought or want to behave, rather than how they actually behave, and artificial intelligence techniques should capture ideal behavior.

Another shortcoming of some artificial intelligence approaches is traceability. Artificial intelligence can be complex, and artificial neural networks specifically are unable to explain in a comprehensible form how a decision was made on the basis of the input data. Already there is anecdotal evidence of computers that have discovered relationships in science that researchers did not understand (*48*). The relationships existed but were incomprehensible to humans, hidden

in gigabytes of data and linkages within an artificial neural network. Bostrom and Yudkowsky have argued that opaque systems cannot be inspected, are unpredictable, and can be manipulated easily (*49*). They recommend the use of decision trees to encourage transparency, which is another type of deontology (Asimov's laws can be formulated easily as a decision flowchart). The risk of manipulation, however, is of particular importance in road vehicle automation. Ethics would likely require that all humans be given equal value, yet a vehicle manufacturer has an incentive to build vehicles that protect its own occupants foremost. Who would buy a car that might protect a stranger at the risk of your or your own family's safety? A self-protection component built into the automated vehicle's ethics could be hidden in a complicated neural network and discoverable only through the analysis of long-term crash trends. Safeguards must be in place to ensure that such a thing does not happen.

Although artificial intelligence approaches allow computers to learn human ethics without the need for humans to perform the difficult task of articulating ethics as code, they produce actions that cannot be justified or explained in an understandable way. If trained with a narrow set of data, an artificial intelligence may learn behaviors that are completely unintended and undesirable. Without further testing, artificial intelligence approaches cannot be recommended for automated vehicles without human-designed rules to increase transparency and prevent obviously unethical behavior.

PROPOSED ETHICAL VEHICLE DEPLOYMENT STRATEGY

This study investigated issues in ethical decision making in automated vehicles from findings in philosophy, artificial intelligence, and robotics. On the basis of the identified strengths and weaknesses of rational and artificial intelligence approaches, the following three-phase approach is proposed, to be enforced as technology becomes available.

Phase 1. Rational Ethics

The first phase, feasible with current technology, would use a rational system for automated vehicle ethics. This system would reward behaviors that minimized global damage. The standards for such a system should be agreed on by developers of automated vehicles, lawyers, transportation engineers, and ethicists, and should be open and transparent to discourage automakers from building self-protection into the algorithms that is excessive. Rules should consist of widely agreed-on concepts (e.g., injuries are preferable to death, property damage is preferable to injury, and vulnerable users should be protected foremost).

A safety metric should be developed for use in situations in which the higher-level rules do not specify a behavior (e.g., when two alternatives each result in similar injuries). This safety metric should be independent of current insurance industry standards, and instead use expertise from ethicists and from existing research. A possible starting point for such a system might include value-of-life estimates, used in medicine, and the identification of organ transplant recipients, in which a complex moral decision must have a numerical basis.

It is unlikely that any human-developed rule set that governs robotic behavior will cover all possible scenarios. In any scenario not covered by the rules, or where the rules conflict or the ethical action is uncertain, the vehicle should brake and evade.

Phase 2. Hybrid Rational and Artificial Intelligence Approach

In the second phase, which requires sophisticated software that does not yet exist, an automated vehicle's software can use machine learning techniques to understand the correct ethical decision, while bound by the rule-based system in Phase 1. A neural network is a likely candidate method for this approach. The neural network would be trained on a combination of simulation and recordings of crashes and near crashes. Humans would score potential actions and results as more or less ethical and would be allowed to score outcomes without the time constraint of an actual crash.

Similar concepts have been promoted. Powers has argued for an adaptive incrementalism in machine ethics, which although it does not specify the technique to use to develop an ethical system, acknowledges that a rational-comprehensive approach is impractical because it restricts progress, given limited knowledge in ethics and computer engineering (*50*). An incremental approach, in which a computer could train itself, would allow progress in automated vehicle research in the absence of a perfect moral system. Wallach and Allen described the limitations of both top-down and bottom-up approaches and recommended a similar hybrid approach (*38*).

Care must be taken to provide the algorithm with a diverse set of training scenarios. If a training set is given, which is too narrow in scope, the algorithm can learn morals that were not intended, similar to an extremist. To ensure reasonable behavior by the vehicle, boundaries should be provided. In Phase 2, the rule system from Phase 1 should remain in place as boundary conditions, and the machine learning approach should focus on scenarios not covered by the Phase 1 rules.

Phase 3. Feedback with Natural Language

Of the available artificial intelligence methods, artificial neural networks are well suited to classify the morality of a vehicle's decision, given their high classification accuracy with large volumes of data with low computational costs. A major shortcoming of a neural network is its incapability to explain its decision. Unlike a decision tree, in which the logic can be traced back over several steps to its source, a neural network is not easily reverse-engineered, and it can be difficult to determine how it arrived at its decision. In an automated vehicle crash, an understanding of the logic behind an automated vehicle's actions is critical, particularly if the vehicle did not behave as expected. The use of onboard data collection may allow researchers to recreate a vehicle's behavior. Even with extensive testing, however, only a probabilistic understanding of its mechanisms is possible in any situation. Without the knowledge of why an automated vehicle behaves a certain way, there is no way to fix the problem to ensure that it will not happen again.

To improve a neural network's comprehensibility, computer scientists have developed techniques to extract rule-based explanations from neural networks that are understandable by a human. This process essentially translates a neural network's internal knowledge into a set of symbolic rules, which can then be expressed as natural language (*51*). Not every decision is accurately represented by a rule, and some rules may be overly complex. Regardless, rule extraction will likely be a useful starting point toward an understanding of the logic of neural networks and, similarly, the decisions of automated vehicles.

Although recent research shows promise (*52*), rule extraction is currently in the research phase. As the science progresses, it should be implemented into automated vehicle ethics systems that use the artificial intelligence techniques from Phase 2.

Summary of Three-Phase Approach

The three phases constitute an incremental approach to automated vehicle ethics. Its best analogy may be the moral education of a child, discussed by Wallach and Allen (*38*). Although a child does not have his full moral ability, and may never reach the highest stage of moral development, parents still enforce behavioral boundaries. Parents also encourage their children to consider ways to think about morals, in the hope that one day they will reach a higher stage. The three-phase approach is essentially an attempt to teach a computer ethics and to ensure that it behaves ethically while it learns.

CONCLUSIONS

Three arguments were made in this paper: automated vehicles will almost certainly crash, even in ideal conditions; an automated vehicle's decisions that precede certain crashes will have a moral component; and there is no obvious way to effectively encode human morality in software.

A three-phase strategy to develop and regulate moral behavior in automated vehicles is proposed, to be implemented as technology progresses. The first phase is a rationalistic moral system for automated vehicles that takes action to minimize the impact of a crash on the basis of generally agreed-on principles (e.g., injuries are preferable to fatalities). The second phase introduces machine learning techniques to study human decisions across a range of real-world and simulated crash scenarios to develop similar values. The rules from the first approach remain in place as behavioral boundaries. The final phase requires an automated vehicle to express its decisions with natural language so that its highly complex logic, potentially incomprehensible to humans, may be understood and corrected.

Researchers have made incredible advances in road vehicle automation, with potentially immense safety benefits. Many of the problems faced by automated vehicles can be overcome with better algorithms or sensors. In contrast, the ethical decision making of automated vehicles requires a vehicle not only to behave ethically but to understand and apply ethics in new situations, even when humans are not in agreement on the ethical choice. Further research into machine ethics is encouraged as it applies to road vehicle automation, particularly with respect to the capability of existing crash mitigation systems to behave ethically in realistic scenarios, the types and frequencies of roadway situations that require ethics, and the legitimacy of a vehicle's tendency to protect its passengers foremost. States that are beginning to legislate vehicle automation should consider not only the general precrash behavior of these vehicles but also the logic and "values" that these vehicles possess.

ACKNOWLEDGMENTS

Special thanks go to Barbara Johnson of the University of Virginia and Patrick Lin of the California Polytechnic State University for their reviews of drafts of this paper.

REFERENCES

1. NHTSA. *Preliminary Statement of Policy Concerning Automated Vehicles.* Publication NHTSA 14-13. U.S. Department of Transportation, Washington, D.C., 2013.
2. Dickmanns, E. D. The Development of Machine Vision for Road Vehicles in the Last Decade. *Proc., IEEE Intelligent Vehicle Symposium,* Versailles, France, Vol. 1, June 2002, pp. 268–281.
3. Jochem, T., D. Pomerleau, B. Kumar, and J. Armstrong. PANS: A Portable Navigation Platform. *Proc., Intelligent Vehicles 1995 Symposium,* Detroit, Mich., 1995.
4. Pomerleau, D. RALPH: Rapidly Adapting Lateral Position Handler. *Proc., Intelligent Vehicles 1995 Symposium,* Detroit, Mich., 1995.
5. Broggi, A., M. Bertozzi, and A. Fascioli. Architectural Issues on Vision-Based Automatic Vehicle Guidance: The Experience of the ARGO Project. *Real-Time Imaging,* Vol. 6, No. 4, 2000, pp. 313–324.
6. *Special Report 253: National Automated Highway System Research Program: A Review.* TRB, National Research Council, Washington, D.C., 1998.
7. Markoff, J. Crashes and Traffic Jams in Military Test of Robotic Vehicles. *New York Times,* Nov. 5, 2007.
8. Markoff, J. Google Cars Drive Themselves, in Traffic. *New York Times,* Oct. 9, 2010.
9. Hachm, M. CES 2013: Audi Demonstrates Its Self-Driving Car. *Popular Science,* Jan. 9, 2013. http://www.popsci.com/cars/article/2013-01/ces-2013-audi-demonstrates-its-self-driving-car. Accessed June 27, 2013.
10. Newcomb, D. Ford Inches Toward Autonomous Cars, Helps the Parking-Challenged. *Wired: Autopia,* June 26, 2012. http://www.wired.com/autopia/2012/06/ford-tech-driving-challenged/. Accessed June 27, 2013.
11. Travers, J. BMW Traffic Jam Assistant Puts Self-Driving Car Closer Than You Think. *Consumer Reports,* June 11, 2013. http://news.consumerreports.org/cars/2013/06/bmw-traffic-jam-assistant-puts-self-driving-car-closer-than-you-think.html. Accessed June 27, 2013.
12. Daimler. The New Mercedes-Benz S-Class Intelligent Drive: Networked with All Senses. Undated. https://www.daimler.com/dccom/0-5-1597521-1-1597533-1-0-0-1597522-0-0-135-0-0-0-0-0-0-0-0.html. Accessed June 27, 2013.
13. Newman, J. Cadillac Has Self-Driving Cars, Too. *Time,* April 20, 2012. http://techland.time.com/2012/04/20/cadillac-has-self-driving-cars-too/. Accessed June 27, 2013.
14. Guizzo, E. Toyota's Semi-Autonomous Car Will Keep You Safe. *IEEE Spectrum Automation,* Jan. 8, 2013. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/toyota-semi-autonomous-lexus-car-will-keep-you-safe. Accessed June 27, 2013.
15. T.S. How Does a Self-Driving Car Work? *Economist,* April 29, 2013. http://www.economist.com/blogs/economist-explains/2013/04/economist-explains-how-self-driving-car-works-driverless. Accessed Aug. 1, 2013.
16. Hyde, J. This Is Google's First Self-Driving Car Crash. *Jalopnik,* Aug. 5, 2011. http://jalopnik.com/5828101/this-is-googles-first-self+driving-car-crash. Accessed June 18, 2013.
17. Yarrow, J. Human Driver Crashes Google's Self Driving Car. *Business Insider,* Aug. 5, 2011. http://www.businessinsider.com/googles-self-driving-cars-get-in-their-first-accident-2011-8. Accessed June 19, 2013.
18. Bilger, B. Auto Correct: Has the Self-Driving Car at Last Arrived? *New Yorker,* Nov. 25, 2013.
19. Smith, B. W. *Driving at Perfection.* The Center for Internet and Society at Stanford Law School, March 2012. http://cyberlaw.stanford.edu/blog/2012/03/driving-perfection. Accessed Oct. 3, 2012.
20. U.S. Census Bureau. *Vehicles Involved in Crashes by Vehicle Type, Rollover Occurrence, and Crash Severity: 2009.* Statistical Abstract of the United States. Publication Table 1107. U.S. Department of Commerce, 2012.
21. Office of Freight Management and Operations. *Freight Facts and Figures 2011.* Publication FHWA-HOP-12-002. FHWA, U.S. Department of Transportation, 2011.
22. NHTSA. *Traffic Safety Facts 2009: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System.* Publication DOT HS 811 402. U.S. Department of Transportation, 2009.
23. Laugier, C., and R. Chatila (eds.). *Autonomous Navigation in Dynamic Environments.* Springer-Verlag, Berlin, 2007.
24. Fraichard, T., and H. Asama. Inevitable Collision States: A Step Towards Safer Robots? *Advanced Robotics,* Vol. 18, No. 10, 2004, pp. 1001–1024.
25. Benenson, R., T. Fraichard, and M. Parent. Achievable Safety of Driverless Ground Vehicles. *Proc., 10th International Conference on Control, Automation, Robotics and Vision,* Hong Kong, 2008, pp. 515–521.

26. Bautin, A., L. Martinez-Gomez, and T. Fraichard. Inevitable Collision States: A Probabilistic Perspective. *Proc., IEEE International Conference on Robotics and Automation,* Anchorage, Alaska, 2010, pp. 4022–4027.
27. Rumar, K. The Role of Perceptual and Cognitive Filters in Observed Behavior. In *Human Behavior and Traffic Safety* (L. Evans and R. C. Schwing, eds.), Plenum Press, New York, 1985, pp. 151–170.
28. Llaneras, R. E., J. A. Salinger, and C. A. Green. Human Factors Issues Associated with Limited Ability Autonomous Driving Systems: Drivers' Allocation of Visual Attention to the Forward Roadway. *Proc., 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design,* Bolton Landing, N.Y., 2013.
29. AASHTO. *A Policy on Geometric Design of Highways and Streets,* 6th ed. Washington, D.C., 2011.
30. Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. *The 100-Car Naturalistic Driving Study, Phase II. Results of the 100-Car Field Experiment.* Publication DOT HS 810 593. Virginia Tech Transportation Institute, Virginia Polytechnic Institute and State University, Blacksburg, 2006.
31. Goodall, N. J. Machine Ethics and Automated Vehicles. In *Road Vehicle Automation* (G. Meyer and S. A. Beiker, eds.), Springer, Berlin, 2014.
32. Marcus, G. Moral Machines. *New Yorker* Blogs, Nov. 27, 2012. http://www.newyorker.com/online/blogs/newsdesk/2012/11/google-driverless-car-morality.html. Accessed March 8, 2013.
33. Sherwood, C. P., J. M. Nolan, and D. S. Zuby. Characteristics of Small Overlap Crashes. *Proc., 21st International Technical Conference on the Enhanced Safety of Vehicles,* Stuttgart, Germany, 2009.
34. Arkin, R. C. Governing Lethal Behavior in Autonomous Robots. CRC Press, Boca Raton, Fla., 2009.
35. Finn, A., and S. Scheding. *Developments and Challenges for Autonomous Unmanned Vehicles: A Compendium.* Springer, Berlin, 2010.
36. Meuhlhauser, L., and L. Helm. Intelligence Explosion and Machine Ethics. In *Singularity Hypotheses: A Scientific and Philosophical Assessment* (A. H. Eden, J. H Moor, J. H. Soraker, and E. Steinhart, eds.), Springer, Berlin, 2012, pp. 101–126.
37. Asimov, I. Runaround. *Astounding Science Fiction.* March 1942, pp. 94–103.
38. Wallach, W., and C. Allen. *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, New York, 2009.
39. Beavers, A. F. Moral Machines and the Threat of Ethical Nihilism. In *Robot Ethics: The Ethical and Social Implication on Robotics,* MIT Press, Cambridge, Mass., 2011.
40. Summers, S., A. Prasad, and W. T. Hollowell. NHTSA's Vehicle Compatibility Research Program. Publication 1999-01-0071. SAE International, Warrendale, Pa., 1999.
41. Evans, L. Death in Traffic: Why Are the Ethical Issues Ignored? *Studies in Ethics, Law, and Technology,* Vol. 2, No. 1, 2008.
42. Evans, L. *Traffic Safety.* Science Serving Society, Bloomfield, Mich., 2004.
43. Evans, L. Causal Influence of Car Mass and Size on Driver Fatality Risk. *American Journal of Public Health,* Vol. 91, No. 7, 2001, pp. 1076–1081.
44. Russell, S. J., and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Upper Saddle River, N.J., 2010.
45. Guarini, M. Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems,* Vol. 21, No. 4, 2006, pp. 22–28.
46. Hibbard, B. Avoiding Unintended AI Behaviors. *Proc., 5th International Conference Artificial General Intelligence,* Oxford, United Kingdom, 2012.
47. Batavia, P., D. Pomerleau, and C. Thorpe. *Applying Advanced Learning Algorithms to ALVINN.* Publication CMU-RI-TR-96-31. Robotics Institute, Carnegie Mellon University, Pittsburgh, Pa., Oct. 1996.
48. Arbesman, S. Explain It to Me Again, Computer. *Slate,* Feb. 25, 2013. http://www.slate.com/articles/technology/future_tense/2013/02/will_computers_eventually_make_scientific_discoveries_we_can_t_comprehend.single.html. Accessed Feb. 25, 2013.
49. Bostrom, N., and E. Yudkowsky. The Ethics of Artificial Intelligence. In *Cambridge Handbook of Artificial Intelligence* (K. Frankish and W. M. Ramsey, eds.), Cambridge University Press, United Kingdom, 2013.
50. Powers, T. M. Incremental Machine Ethics. *IEEE Robotics Automation Magazine,* Vol. 18, No. 1, 2011, pp. 51–58.
51. Tickle, A. B., R. Andrews, M. Golea, and J. Diederich. The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks. *IEEE Transactions on Neural Networks,* Vol. 9, No. 6, 1998, pp. 1057–1068.
52. Augasta, M. G., and T. Kathirvalavakumar. Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters,* Vol. 35, No. 2, April 2012, pp. 131–150.

The Vehicle–Highway Automation Committee peer-reviewed this paper.