MORAL CRUMPLE ZONES

When your self-driving car crashes, you could still be the one who gets sued

Madeleine Clare Elish & Tim Hwang | July 25, 2015

Just because there's a steering wheel somewhere doesn't mean you're in control. (Reuters/Arnd Wiegmann)

In 1969, Milton Packin was pulled over for speeding on a New Jersey highway. He appealed the ticket, claiming that he wasn't driving the car; it was the cruise

no[t] less responsible for its operation," he stated. Packin may not have been directly in control of the gas being fed to the motor, but he was still responsible for the speed of the car.

in today's Daily Brief.

But what about when cars are completely self-driving? If you delegate transportation from point A to Point B entirely to a machine, are you responsible if it hits someone? As current laws stand, you probably would be—but should you be?

Traditionally, the more control you have over something, the more responsible you are. The Packin case was straightforward in part because control and responsibility seemed proportionate to one another: Packin had the power to turn on the cruise control as well as to set the speed. It made sense that he was responsible for its behavior. With increasingly complex automation and autonomous technologies, things will not be as clear-cut.

In a self-driving car, the control of the vehicle is shared between the driver and the car's software. How the software behaves is in turn controlled—designed—by the software engineers. It's no longer true to say that the driver is in full control, as the judge declared in the Packin case. Nor does it feel right to say that the software designers are entirely control.

Yet as *control* becomes distributed across multiple actors, our social and legal conceptions of *responsibility* are still generally about an individual. If there's a crash, we intuitively—and our laws, in practice—want some*one* to take the blame.

The result of this ambiguity is that humans may emerge as "liability sponges" or "moral crumple zones." Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in an autonomous system may become simply a component—accidentally or intentionally—that is intended to bear the brunt of the moral and legal penalties when the overall system fails.

So how do we update our laws and social norms about responsibility to this new reality of control?

To understand the roads, look to the skies

Debris from AF 447, recovered from the Atlantic Ocean. (Reuters/JC Imagem/Alexandre Severo)

Aviation is at the vanguard of complex and highly automated human-machine systems, and might provide a hint of what awaits us on the roads. A modern aircraft spends most of its time in the air under the control of a set of technologies, including an autopilot, GPS and flight management system, which govern almost everything it does.

Yet even while the plane is being run by software, the pilots in the cockpit are legally responsible for its operation. US Federal Aviation Administration (FAA) regulations specify this directly, and courts have consistently upheld it. So when something goes wrong, we observe pilots becoming "moral crumple zones"—largely totemic humans whose central role becomes soaking up fault, even if they had only partial control of the system.

The classic example is the case of Air France flight 447. En route from Brazil to France in 2009, the Airbus 330 had flown into a storm, and ice crystals had formed on the plane's pitot tubes, a part of the avionics system that measures air speed. The frozen pitot tubes sent faulty data to the autopilot. The autopilot did what it was designed to do in the absence of data: It automatically disengaged, bouncing control of the aircraft back to the pilots. Caught by surprise, they were overwhelmed by an influx of loud warning signals and confusing instrument readings. In the words of the official French report, they lost "cognitive control of the situation." A series of errors and incorrect maneuvers by the pilots ended in the plane's crashing into the Atlantic Ocean, killing all 228 people on board.

News reports at the time emphasized this series of pilot errors. Almost as a footnote, they explained that there were other factors involved, including the known but not yet fixed mechanical problem with pitot tubes failing due to icing in A330s. What contemporary reports did not point out—though a later essay in Vanity Fair brought to public attention—was that the pilots' mistakes, some of them rookie errors, may have been at least partly due to automation. They had to take charge of the aircraft under conditions of weather and altitude that modern pilots rarely experience, because the autopilot is almost always in control.

As pilots fly less and "supervise" automatic systems more, their basic flying skills can atrophy—the so-called "automation paradox." As Earl Wiener, a leader in the field of human factors in aviation, prophesied, automation has not eliminated

human error so much as created opportunities for new kinds of error, born out of miscommunication between humans and machines.

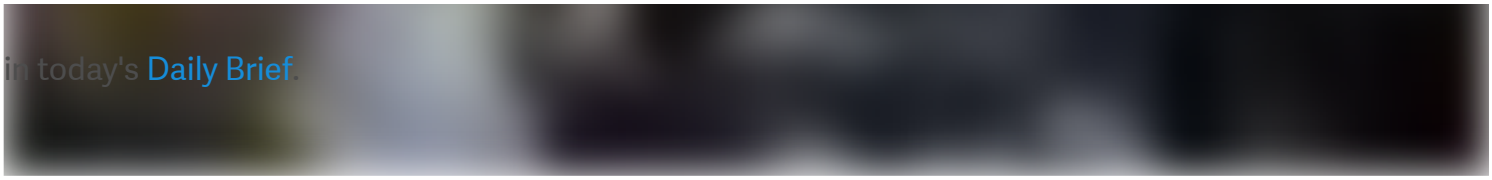
in today's Daily Brief.

Part of the problem is that cultural perceptions of automation tend to elevate automation and criticize humans (pdf). For instance, when the A330's predecessor, the A320, was unveiled in 1986, an aviation expert was quoted as saying that the plane's all-new fly-by-wire system would "be smart enough to protect the airplane and the people aboard it from any dumb moves by the pilot." In the case of AF 447, automation seems, instead, to have contributed to those "dumb moves."

This is not to say automation is bad *per se*. Commercial aviation has become much safer over the years, and experts agree that increasing automation is largely responsible. It is expected to similarly transform safety on the roads. Still, contemporary culture has developed a schizophrenic attitude toward automation: It is our salvation, freeing us from dull, dirty and dangerous jobs, but it is also our nightmare, resulting in the mental and physical evisceration of our species.

The false reassurance of a "human in the loop"

in today's Daily Brief.



"Are you sure you want to fly to Kabul? This action cannot be undone." (Reuters/Sergio Perez)

In part, designers and engineers have met this anxiety by assuring us that a "human is in the loop." Emerging from a World War II cybernetic lineage, human-in-the-loop design principles require that an automated action always include some sort of human input, even if that input is just clicking "OK." Tesla has announced that in its soon-to-be-launched semiautonomous models, the driver will be required to press a turn signal button in order for the car to complete a planned maneuver, like turning a corner or passing another car. The car will plan and do the driving, but a human will have to sign off on each move.

The human-in-the-loop principle may be well intentioned—after all, it's meant to theoretically ensure that human judgment can trump automation. But as we've seen in the context of aviation, it means the human takes on a responsibility disproportionate to the amount of real-time control they are allowed. And without the practice of hands-on driving, the human is likely to be slow at reacting when something *does* go wrong.

The risk, then, is that human-in-the-loop design principles become not a way for human drivers to retain control so much as a way for system designers to deflect responsibility and to create "moral crumple zones." As long as responsibility can fall only on one side or the other, the designers have an incentive to build systems, like a driverless car, that give a driver just enough control for a court to always rule that any crash is the driver's fault.

There's an intimation of this in early results from Google's self-driving cars. Earlier this summer, Google made public the accident record of car tests from May 2010 to

testing, prepared to take over if anything went wrong. Google declared that none of the accidents had been caused by the car; all were the fault of human drivers.

There was, however, a surprising pattern: 10 of the 12 accidents were rear-end collisions. It's possible that these kinds of accidents are the most common on the streets and highways around Palo Alto. But it's also possible that the Google car effectively contributed to some of the accidents by driving in a way that drivers around it didn't expect—more slowly or cautiously, for instance, than a typical human driver.

In other words, just as AF 447 crashed because of a miscommunication between the pilots and their automated systems, the Google car accidents might have been caused by a fundamental miscommunication between a driverless car and a human-driven car. And though this miscommunication is two-way—the result of the Google car not behaving as expected—it's the human drivers who become the "moral crumple zones," taking on responsibility for a failure when in fact control over the situation was shared.

What's the solution?

So many choices, so little control. (AP Photo/Matt Dunham)

So is the answer to take the human out of the loop—to put the car's software exclusively in control, so it can be blamed if there's a crash? That might solve the aforementioned "automation paradox" —no more expecting a drowsy, unpracticed human driver to make split-second decisions in a crisis—but society might not be ready to trust machines that much. Machines may not be capable of complete control, all the time and everywhere. And in any case, it still leaves the problem of how to hold machines responsible.

Some have suggested that the solution to both problems is to bestow legal personhood upon effectively autonomous agents, so they can be held liable and carry insurance. The idea is not as crazy as it seems—in America, corporations are persons, after all—and it might give victims of accidents quick redress. However, it also creates the potential for the designers of these systems to evade responsibility for technological and design choices that may have contributed to the accident.

What is certain is that operators and consumers of autonomous systems shouldn't be left holding the bag when things go wrong. Despite all the press, autonomous cars are not yet on the roads. Military grade lethal weapons are still under significant human control. Truly autonomous technologies are still in the development phase. Now is the time to think carefully and create laws and standards proactively.

We need to keep in mind the new nature of shared control, potentially dispersed in space and time. We need to think about the context in which assumptions about

responsibility are being made. The question before us is not how to make computational agents responsible, but rather, how to apply responsibility fairly.

We welcome your comments at ideas@qz.com.