ETHICS

Our driverless dilemma

When should your car be willing to kill you?

By **Joshua D. Greene**

Suppose that a driverless car is headed toward five pedestrians. It can stay on course and kill them or swerve into a concrete wall, killing its passenger. On page 1573 of this issue, Bonnefon *et al.* (*1*) explore this social dilemma in a series of clever survey experiments. They show that people generally approve of cars programmed to minimize the total amount of harm, even at the expense of their passengers, but are not enthusiastic about riding in such "utilitarian" cars—that is, autonomous vehicles that are, in certain emergency situations, programmed to sacrifice their passengers for the greater good. Such dilemmas may arise infrequently, but once millions of autonomous vehicles are on the road, the improbable becomes probable, perhaps even inevitable. And even if such cases never arise, autonomous vehicles must be programmed to handle them. How should they be programmed? And who should decide?

Bonnefon *et al.* explore many interesting variations, such as how attitudes change when a family member is on board or when the number of lives to be saved by swerving gets larger. As one might expect, people are even less comfortable with utilitarian sacrifices when family members are on board and somewhat more comfortable when sacrificial swerves save larger numbers of lives. But across all of these variations, the social dilemma remains robust. A major determinant of people's attitudes toward utilitarian cars is whether the question is about utilitarian cars in general or about riding in them oneself.

In light of this consistent finding, the authors consider policy strategies and pitfalls. They note that the best strategy for utilitarian policy-makers may, ironically, be to give up on utilitarian cars. Autonomous vehicles are expected to greatly reduce road fatalities (*2*). If that proves true, and if utilitarian cars are unpopular, then pushing for utilitarian cars may backfire by delaying the adoption of generally safer autonomous vehicles.

Department of Psychology, Center for Brain Science, Harvard University, Cambridge, MA 02138, USA. Email: jgreene@wjh.harvard.edu

ILLUSTRATION: DARIA KIRPACH/@SALZMANART

Moral dilemma. Should autonomous vehicles protect their passengers or minimize the total amount of harm?

As the authors acknowledge, attitudes toward utilitarian cars may change as nations and communities experiment with different policies. People may get used to utilitarian autonomous vehicles, just as some Europeans have grown accustomed to opt-out organ donation programs (*3*) and Australians have grown accustomed to stricter gun laws (*4*). Likewise, attitudes may change as we rethink our transportation systems. Today, cars are beloved personal possessions, and the prospect of being killed by one's own car may feel like a personal betrayal to be avoided at all costs. But as autonomous vehicles take off, car ownership may decline as people tire of paying to own vehicles that stay parked most of the time (*5*). The cars of the future may be interchangeable units within vast transportation systems, like the cars of today's subway trains. As our thinking shifts from personal vehicles to transportation systems, people might prefer systems that maximize overall safety.

In their experiments, Bonnefon *et al.* assume that the autonomous vehicles' emergency algorithms are known and that their expected consequences are transparent. This need not be the case. In fact, the most pressing issue we face with respect to autonomous vehicle ethics may be transparency. Life-and-death trade-offs are unpleasant, and no matter which ethical principles autonomous vehicles adopt, they will be open to compelling criticisms, giving manufacturers little incentive to publicize their operating principles. Manufacturers of utilitarian cars will be criticized for their willingness to kill their own passengers. Manufacturers of cars that privilege their own passengers will be criticized for devaluing the lives of others and their willingness to cause additional deaths. Tasked with satisfying the demands of a morally ambivalent public, the makers and regulators of autonomous vehicles will find themselves in a tight spot.

Software engineers—unlike politicians, philosophers, and opinionated uncles—don't have the luxury of vague abstraction. They can't implore their machines to respect people's rights, to be virtuous, or to seek justice—at least not until we have moral theories or training criteria sufficiently precise to determine exactly which rights people have, what virtue requires, and which trade-offs are just. We can program autonomous vehicles to minimize harm, but that, apparently, is not something with which we are entirely comfortable.

Bonnefon *et al.* show us, in yet another way, how hard it will be to design autonomous machines that comport with our moral sensibilities (*6*–*8*). The problem, it seems, is more philosophical than technical. Before we can put our values into machines, we have to figure out how to make our values clear and consistent. For 21st-century moral philosophers, this may be where the rubber meets the road. ■

REFERENCES

1. J.-F. Bonnefon *et al.*, *Science* **352**, 1573 (2016).
2. P. Gao, R. Hensley, A. Zielke, *A Road Map to the Future for the Auto Industry* (McKinsey & Co., Washington, DC, 2014).
3. E. J. Johnson, D. G. Goldstein, *Science* **302**, 1338 (2003).
4. S. Chapman *et al.*, *Injury Prev.* **12**, 365 (2006).
5. D. Neil, "Could self-driving cars spell the end of car ownership?", *Wall Street Journal*, 1 December 2015; www.wsj.com/articles/could-self-driving-cars-spell-the-end-of-ownership-1448986572.
6. I. Asimov, *I, Robot* [stories] (Gnome, New York, 1950).
7. W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford Univ. Press, 2010).
8. P. Lin, K. Abney, G. A. Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2011).

10.126/science.aaf9534

IMMUNOLOGY

Converting to adapt

Gut microbiota affect T cell plasticity in the intestinal lining

By **Marco Colonna** *and* **Luisa Cervantes-Barragan**

Effective immune responses rely on balancing lymphocyte stability and plasticity. Lymphocytes have regulatory circuits that control phenotypic and functional identity. Stable circuits maintain homeostasis and prevent autoimmunity. But plasticity is needed to integrate new environmental inputs and generate immune responses that subdue the eliciting agent without damaging tissue. Regulatory T cells (T_{regs}) are a subset of $CD4^+$ T cells that control effector T cell responses and prevent excessive inflammation and autoimmunity (*1*, *2*). On page 1581 in this issue, Sujino *et al.* (*3*) report that intestinal T_{regs} convert into $CD4^+$ intraepithelial T cells ($CD4_{IELs}$) to adapt to the local intestinal environment, thus identifying the intestinal epithelium as a compartment that enforces lymphocyte plasticity.

"...Foxp3⁺ cells might rapidly convert into another T cell subtype."

$CD4_{IELs}$ are implicated in various immune responses, including tolerance to dietary antigens (*4*). They originate from $CD4^+$ T helper cells in the intestinal lamina propria, and can produce interferon-γ (IFN-γ), a cytokine that triggers immune responses to infection, as well as promote cytolysis. Differentiation of T cells into $CD4_{IELs}$ is governed by the reduced expression of ThPOK (T helper–inducing POZ/Kruppel factor), a transcription factor that drives the $CD4^+$ T helper cell program. Moreover, increased expression of Runx3 (runt-related transcription factor 3) drives the $CD8^+$ T cell program, i.e. IFN-γ production and cytolysis (*5*, *6*). $CD4_{IELs}$ in the intestinal

Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. Email: mcolonna@pathology.wustl.edu

Our driverless dilemma
Joshua D. Greene (June 23, 2016)
Science **352** (6293), 1514-1515. [doi: 10.1126/science.aaf9534]