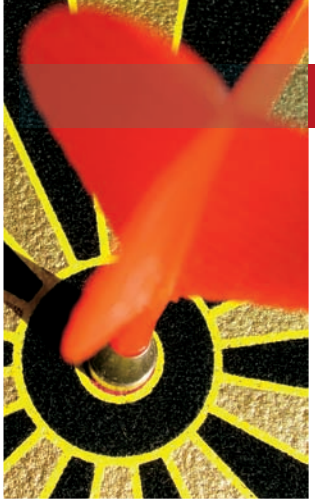

Editor: **Daniel Zeng**, University of Arizona and Chinese Academy of Sciences, zengdaniel@gmail.com

Regulating Autonomous Systems: Beyond Standards

David Danks and Alex John London, *Carnegie Mellon University*

Calls to regulate autonomous systems are escalating, spurred partly by recent adverse events involving self-driving cars.[1,2] For example, the US Department of Transportation (DoT) recently issued a report providing guidance about the development and deployment of autonomous vehicles[3] that will presumably serve as a precursor to some future regulation. However, we contend that autonomous systems pose a fundamental challenge to any traditional regulatory scheme that relies on the promulgation and enforcement of safety and reliability standards. These standards-based regulatory systems presuppose that the regulated products execute only well-defined functions in known, stable contexts for which performance benchmarks can be defined and assessed. This approach is generally adequate for the regulation of *automatic systems*, or systems that perform a delineated set of operations within a well-defined and relatively static context to achieve specific goals.

In contrast, *autonomous systems* move beyond mere automation, as they make meaningful decisions about which tasks to perform in uncertain or ambiguous contexts. Uncertainty about context can arise for multiple reasons, including changes over time or ambiguous signals about key indicators. Regardless of the source of the uncertainty, a necessary condition for autonomy is the ability to identify contexts in a fluid environment and then select and execute appropriate functions in ways that conform to relevant and potentially context-specific norms, constraints, or desiderata. However, this ability creates a prima facie challenge to any attempt to develop performance standards for autonomous systems, because standards presuppose known contexts. Recognizing this challenge is a necessary step toward evaluating alternative models of oversight for novel autonomous systems.

Contextual Certainty, Standards, and Automation

Standardization plays a key role in facilitating technological development by promoting modularity and ensuring uniformity and reliability. We focus in this article on performance (that is, function-centered) rather than design (that is, construction-centered) standards.[4] Autonomous systems are sufficiently complex, and can be realized in enough different ways, that regulations cannot focus purely on the system's design; we care about how they perform—perhaps in a wide range of contexts—and not about how they happen to be built. At a general level, a performance standard is defined by three dimensions critical to the safe, reliable, and appropriate operation of a technology: function, tolerance, and contexts of operation. *Function* specifies what the technology does in order to achieve the goal it is designed to advance. *Tolerance* specifies how closely (or frequently) the technology must achieve the function. *Appropriate contexts of operation* are the environments within which a technology is designed to operate so that it can achieve the intended function to within the required tolerance. This last dimension—appropriate contexts—determines the cases in which a system could correctly be used, whereas the other two determine what the system should do in those (appropriate) contexts of operation.

The function, tolerance, and contexts for a standard are often determined by the technology's role in some higher-order process or system. In particular, performance standards greatly facilitate the development of modular technologies that nest

1541-1672/17/$33.00 © 2017 IEEE
Published by the IEEE Computer Society

together to form more complex systems. In these cases, the contexts of operation are determined partly by the higher-order system, because it can provide clear signals about which function to perform under specific circumstances. A piston, for example, performs a specific function within an engine, and so the particular type of engine—both its design and performance—constitutes the relevant context of operation and helps determine the piston's tolerances. Similarly, the type of vehicle in which an engine will be used constitutes the context for the engine's function and helps determine its tolerances.

The appropriate operation of automatic systems can be specified using performance standards: they build in significant assumptions about their environments, and thus have clear contexts of operation. Automatic systems are often designed to function in specific contexts that are fixed, or that vary in well-defined and easily detectable ways. In some cases, the clear context results from the environment being regularized. For instance, gear shifting can be automated because the engine context (speed and RPM) is sufficiently constrained and regular, such that the shifting function can be based on a fixed set of clear signals. In other cases, automatic systems rely on users to ensure that the context is appropriate. For example, early versions of cruise control automated the process of maintaining a uniform speed, but they relied on the user to ensure that the context was appropriate. The driver still had to steer the car, monitor the vehicle's distance from other vehicles or objects, and decide when to engage, change speed, or disengage cruise control. The scope of these appropriate contexts has expanded as sensor technologies have improved; for example, adaptive cruise control can modulate speed relative to other objects, whether fixed or moving. Even so, such systems continue to perform clear responses to well-defined events with a strong reliance on the driver to detect and respond to relevant changes in the context of operation.

Autonomy and Uncertain Contexts

In contrast, although particular components or modules within autonomous systems can be subject to performance standards, autonomous systems as a whole pose a fundamental challenge for such regulatory standards, precisely because they go beyond the execution of a limited set of functions in a well-defined context. There are many characterizations of "autonomy,"[3,5,6] but they all require that the system be able to determine the context in which it is operating, and thereby modulate its behavior according to context-relevant norms, constraints, or desiderata. Autonomy requires some ability to function independently of human users, in both executive and decision-making functions. In many cases, autonomous systems are desirable precisely because of this independence. For example, one attraction of autonomous vehicles over those with adaptive cruise control is that the latter require a driver to be present to monitor the vehicle performance and intervene if contextual factors change. The automatic system (adaptive cruise control) can perform a narrow class of functions in a limited environment—thus freeing the user from having to perform those functions—but still requires the presence of a user who can intervene once contexts shift.

In contrast, autonomous systems are supposed to operate with greater independence from human users, but this requires that they assume the tasks of discriminating relevant contexts and altering their behavior accordingly. The scope of a system's autonomy is thereby a function of the range of contexts between which it can be trusted to discriminate and appropriately respond. For example, an autonomous truck might be able to haul freight from one location to another without the presence of a human driver, but only if there is a range of contexts within which it can appropriately choose and (safely) navigate an efficient route.

An autonomous military aircraft that can launch, navigate, target, strike, and return would have an even wider scope of autonomy, because it would have to make a wider range of context-sensitive decisions traditionally reserved for the human operator. As with the truck, the autonomous aircraft would have to safely navigate a route to and from the target. Unlike the truck, it would also need to distinguish civilians, friendly forces, and enemy combatants on the basis of noisy inputs, including people's movements. Identifying behavior as benign or hostile requires complex inferences from contextual features that could be highly variable. Appropriate responses given those inferences are even more context-sensitive, because details of the case and context can readily change what constitutes advancing legitimate military goals while respecting the rights and welfare of noncombatants.

Moreover, autonomous systems must be able to identify the contexts even when the environment is noisy, ambiguous, and rapidly changing. In contrast, if the contexts are known in advance, or provide highly reliable signals when they change, an automatic system will suffice. And, in addition to the difficulty of identifying the noisy, variable context, an autonomous system must address significant challenges in identifying

suitable behaviors: its goals may be realizable by multiple actions, each of which must be sensitive to contextually relevant norms. For example, when avoiding an unusual obstacle in the road, a vehicle must also avoid colliding with other vehicles or pedestrians; in targeting enemy combatants, an autonomous weapons system must be able to avoid inflicting harm on noncombatants and friendly forces.

The importance of contextual uncertainty and response selection for autonomous systems can also be seen in canonical worries about those systems going awry. The possibility of autonomous weapons systems indiscriminately targeting civilians often turns on the presence of ambiguous signals (for example, caravans of civilians being mistaken for formations of soldiers).[7] Problematic hypothetical cases of autonomous driving often involve contextual shifts (such as changes in others' driving patterns due to weather, road construction, or other factors).[8]

Regulating Autonomous Systems

The need for regulation has become increasingly urgent as autonomous systems advance in abilities and usage, particularly in the domains of transportation, personal care, and warfare. Performance standards can provide meaningful assurances of safety and reliability when contexts, functions, and tolerances are clear, but these are precisely the cases for which autonomy is not necessary. Autonomous systems must determine appropriate contexts and functions, and so performance standards cannot be specified in advance with the required precision. At best, standards for autonomous systems can merely codify the desire that such systems be reliable (in some sense).

There seems to be little recognition of this mismatch between autonomous systems and traditional regulatory systems based on performance standards. For example, the recent DoT guidance on autonomous vehicles says that it "will continue to exercise its available regulatory authority" and focuses on the DoT's ability "to recall vehicles or equipment that pose an unreasonable risk to safety."[3] This approach fits squarely into the mold of traditional performance standards: define metrics for acceptable performance, check to see if the technology satisfies those criteria, allow it if it does, and ban or recall it if not. Moreover, although the DoT recognizes that changes might be required, it nonetheless suggests only "conducting research to develop and validate new performance *metrics*, establishing minimum or maximum *thresholds* for those *metrics*, developing *test* procedures and *test* equipment, and then … [incorporating] those metrics, procedures, and tests in new FMVSS [standards]" (emphasis added).[3] The current regulations, guidance, and possible future regulations all focus on the same traditional types of performance standards and regulatory system.

In contrast, we suggest two possible regulatory alternatives to performance standards for autonomous systems in general (not just autonomous vehicles). First, we can seek to limit the scope of autonomy in these systems so that we can employ existing regulatory structures. One way to achieve this end is to require inclusion of the best context-recognition device we have: humans. The responsibility for context clarification and recognition would remain with the human, and so the scope of the system's autonomy would be constrained. Such systems can be regulated using performance standards because the human ensures that the context is appropriate, although many potential drawbacks exist to pairing human operators with semiautonomous systems (for example, decreased vigilance and lowered situational awareness[9]). This strategy is currently used with many semiautonomous weapons systems and driving technologies (although not necessarily as explicit regulation). Alternately, the scope of autonomy could be constrained by regularizing the environment. This strategy has explicitly been pursued for aerial vehicles (for example, air traffic control systems), and some terrestrial vehicle contexts (such as limited access points and animal fencing for high-speed roads). Such regularization obviates the need for the system to perform context clarification, and so automation is sufficient and performance standards can be imposed. The key drawback of limiting the scope of autonomy in either of these ways—requiring humans or regularizing contexts—is that we fail to gain the promised benefits of autonomy, because it is either blocked or unnecessary.

Second, and more interestingly, we can shift from the simple, binary view of standards to a staged, dynamic system that resembles the regulatory and approval process for drugs and medical devices.[10] In this model, regulation of autonomous systems would be carried out in phases. The autonomous systems analogue to preclinical evaluation involve testing in simulated environments that are continuously updated to include novel or unforeseen situations that occur in the real world; for example, an autonomous vehicle would be tested in diverse landscapes, climate conditions, and environmental factors. A key goal in this stage is to generate information about how the autonomous system, given its specific array of hardware

and software, generates and uses information to individuate relevant contexts in different environments, select appropriate functions, and conform to operative norms or constraints. Acceptable performance in such simulated environments is followed by the analogue of first-in-human studies: limited introduction into targeted real-world settings by specially trained users capable of monitoring system performance in the wild. A key goal of such early-phase trials is to validate real-world performance in different settings and determine ways in which hardware and software can be altered to improve context recognition and appropriate function choice, and thereby to improve performance.

Successful trials in targeted environments would be required before regulators permit testing in different settings. Just as pharmaceuticals are dispensed by prescription to ensure that they are taken for the correct indication at the right dose and schedule, initial roll-out of autonomous systems should be limited to users who receive special training to identify environments in which systems have been tested, what might constitute a "new" environment for a particular system, and how to monitor systems in such contexts. Restrictions on market access can be further relaxed over time, as systems are refined, problems addressed, and long-term patterns of reliability verified under real-world conditions.

This dynamic approach requires continuous gathering of system data and, more importantly, a regulatory body (analogous to the Food and Drug Administration) that is empowered to review all system performance data for any autonomous system proposed for the market. Such a body or agency would ensure that autonomous systems are evaluated under realistic conditions, rather than in simulated or real settings that are configured to guarantee success. The regulatory body would also work to establish socially agreed-upon "population-level" or outcome-oriented benchmarks for acceptable performance for such systems. In the case of autonomous vehicles, for example, this could involve determining socially acceptable numbers of accidents per year of driving, perhaps adjusted by severity and causes of such incidents.

Like the current system of drug development, this approach has the moral downside that product evaluation relies heavily on detection and response to significant adverse events. But unless contextual uncertainty can be eliminated for such systems, this approach represents the only way to deploy truly autonomous systems, monitor and evaluate their performance, and ensure that their development is responsive to real-world challenges.

Performance standards can ensure the safe and reliable function of some components of autonomous systems, but regulation of the full system requires an approach that can accommodate uncertain and changing contexts while providing data-driven assurance of safety and reliability. A regulatory system analogous to ones already in place for pharmaceuticals and other medical interventions represents an attractive and feasible alternative. ▪

Acknowledgments

The authors are listed in alphabetical order; both contributed equally to this article.

References

1. B. Vlasic and N.E. Boudette, "'Self-Driving Tesla Was Involved in Fatal Crash,' U.S. Says," *New York Times*, 30 June 2016.
2. A. Davies, "Google's Self-Driving Car Caused Its First Crash," *Wired*, 29 Feb. 2016.
3. *Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety*, NHTSA, US Dept. Transportation, 2016.
4. C. Coglianese, J. Nash, and T. Olmstead, "Performance-Based Regulation: Prospects and Limitations in Health, Safety, and Environmental Protection," *Administrative Law Rev.*, vol. 55, no. 4, 2003, pp. 705–729.
5. R. Sparrow, "Killer Robots," *J. Applied Philosophy*, vol. 24, no. 1, 2007, pp. 62–77.
6. M. Maurer et al., eds., *Autonomous Driving: Technical, Legal and Social Aspects*, Springer Open, 2015.
7. P. Scharre, *Autonomous Weapons and Operational Risk*, Center for a New American Security, 2016.
8. J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science*, vol. 352, no. 6293, 2016, pp. 1573–1576.
9. M.R. Endsley and E.O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors*, vol. 37, no. 2, 1995, pp. 381–394.
10. J. Kimmelman and A.J. London, "The Structure of Clinical Translation: Efficiency, Information and Ethics," *Hastings Center Report*, vol. 45, no. 2, 2015, pp. 27–39.

David Danks is the L.L. Thurstone Professor of Philosophy and Psychology, and the head of philosophy, at Carnegie Mellon University. Contact him at ddanks@cmu.edu.

Alex John London is the director of the Center for Ethics and Policy and a professor of philosophy at Carnegie Mellon University. Contact him at ajlondon@andrew.cmu.edu.