The 2016 AAAI Fall Symposium Series:
Cognitive Assistance in Government and Public Sector Applications
Technical Report FS-16-02

Responding to Challenges in the Design of Moral Autonomous Vehicles

Helen Zhao
Johns Hopkins University
hzhao16@jhu.edu

Kirsten Dimovitz, Brooke Staveland, Larry Medsker
The George Washington University
lrm@gwu.edu

Abstract

One major example of promising 'smart' technology in the public sector is the autonomous vehicle (AV). AVs are expected to yield numerous social benefits, such as increasing traffic efficiency, decreasing pollution, and decreasing traffic accidents by 90%. However, a recent 2016 study published by Bonnefon *et al.* argued that manufacturers and regulators face a major design challenge of balancing competing public preferences: a moral preference for "utilitarian" algorithms; a consumer preference for vehicles that prioritize passenger safety; and a policy preference for minimum government regulation of vehicle algorithm design. Our paper responds to the 2016 study, calling into question the importance of explicitly moral algorithms and the seriousness of the challenge identified by Bonnefon *et al.* We conclude that the 'social dilemma' is probably overstated. Given that attempts to resolve the 'social dilemma' are likely to delay the rollout of socially beneficial AVs, we implore the need for further research validating Bonnefon *et al.*'s conclusions and encourage manufacturers and regulators to commercialize AVs as soon as possible. We discuss the implications of this example for AV's for the larger context of Cognitive Assistance in other application areas and the government and public policies that are being discussed.

Introduction

Technology has advanced rapidly for Cognitive Assistance systems that are human-friendly and provably safe. As decision-making increasingly falls to artificial intelligence, scientists and system developers need to be in conversation with disciplinary experts from philosophy to policymaking to anticipate the impact of Cognitive Assistance on society and what if any public policy measures should be taken. This calls for cross-disciplinary research and collaboration to anticipate problems and recommend policies for the future, including implications with regard to the regulatory role of government.

The larger context of this paper is the moral dilemma associated with Cognitive Assistance, which seeks to improve performance on complex tasks in which people and machines are treated as complementary co-systems working together. Our case study exemplifies the cross-disciplinary collaboration typical of artificial intelligence research, and focuses on the important and imminent case of 'smart' technology in the public sector: the autonomous vehicle (AV).

Recent advances in artificial intelligence (AI) such as the development and use of human-computer systems to augment human capabilities and replace the need for some human activities raise questions about both what it means to be human and how the use of cognitive assistance should be regarded. Increasing support and consumer demand for smart systems and products foster an escalating pace AI-based product development. This technical advancement frequently gets ahead of the need for the need to discuss implications and corresponding public regulation and private restraint within the social sphere. Cultural values and norms, including what it means to be human and the role of autonomous entities, impact public and private affairs, including law, politics, and social conduct. This paper aims to encourage multi-disciplinary scholars to focus on the realistic potential of AI and what reaction public and private institutions should do regarding the influence on the human condition.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The Dilemma of Cognitive Assistance

The development of autonomous weapons systems is an important example of dilemmas arising about the balance of the human and machine collaboration, including whether cognitive assistance can and should lead to a completely autonomous system. Ronald Arkin [1] addresses the status and future of autonomous weapons systems, the potential benefits and dangers, and the legal and ethical implications of autonomous AI systems inside and outside the military areas. He considers the many issues surrounding the use of lethal autonomous weapons systems from a variety of legal, ethical, operational, and technical perspectives. International discussions and conferences broach the idea that if these systems are developed appropriately they may have the ability to reduce civilian casualties significantly in the battle space. This could lead to a moral imperative for their use, not unlike the use of precision-guided munitions in urban settings to reduce noncombatant deaths.

Healthcare data and systems are arguably the most significant issues of our time. Follow on projects from IBM's project Watson in the Journey to Cognitive Computing program (Nahomoo 2014) are in the important space of Big Data driven by unstructured data comprising video, image, audio, text, and structured data. IBM is ushering in a new era of computing, transitioning from tabulating systems to ubiquitous programmable systems that of the cognitive assistance era with all the issues and moral dilemmas deriving from the blend humans and machines and the impact on individuals and society. The associated research and development of AI systems using deep learning attempts to equip autonomous systems that mimic human capabilities to deal with deeper evidence so that the cognitive assistant can employ deep content analysis, natural language processing, information extraction, and deep machine learning. Google's DeepMind shows promise for learning from humans and going beyond our best experts autonomously. Thus, cognitive enhancement systems might improve human expert performance as well as help retain cognitive functioning as humans age.

Concerns regarding legal implications and governmental regulatory role of the implicit participation of machines in human decision making are not new. The issue has become salient with the realization that the roles of humans and machines in autonomous systems are becoming sufficiently interdependent that they are inseparable and present a challenge to legal frameworks for assessing guilt and liability. Because the intent to use a cognitive utensil is a voluntary and conscious choice of the operator (or decision-maker), the operator retains responsibility and liability for consequences of its use when it functions correctly. Joint human-machine cognitive systems may differ essentially when they use unobtrusively sensed information about a machine operator's cognitive state to overcome transient bottlenecks in the operator's available cognitive resources. Hence, the interaction is implicit because the augmentation occurs automatically, without conscious activity of the human components. These concepts define the parameters of the responsibility of decision makers for their actions, so the actions of advanced human-machine cognitive technologies must be assessed within that realm of discourse. The relevant philosophical and legal concepts, including the taxonomy of excuses, provide a framework for exploring the status of the human components at two levels (1) as designer (and programmer) of machines and automatic processes and (2) as active, real-time participants in the acts of human-machine cognitive systems. (Balaban 2014).

The commonality of moral dilemmas arising from the advancements in cognitive assistance and autonomous vehicles challenge our legal, moral, and policy-making systems. We use the analysis of questions on AV's to look for common questions and understandings for other cognitive assistance

Cognitive Assistance Example: Design of Moral Autonomous Vehicles

(Bonnefon, Shariff, and Rahwan 2016) recently published a study evaluating people's attitudes towards AVs. The study surveyed a total of 1,928 participants on their preferences for AV programming in collision scenarios where harm is unavoidably inflicted on pedestrians or passengers. It presented participants with various cases, such as sacrificing 1 passenger to save 10 pedestrians; sacrificing 1 pedestrian to save 1, 20, or 100 pedestrians; and sacrificing 2 passengers, one of which is a family member, to save 20 pedestrians. It was observed that most people considered "utilitarian" AVs—vehicles programmed to sacrifice a few lives to save more lives—to be morally preferable. Nevertheless, a majority also disfavored purchasing utilitarian AVs for themselves due to the threat of personal sacrifice, and disapproved of government regulation enforcing the design of utilitarian AVs.

Based on their results, the authors concluded that manufacturers and regulators of ethical AVs face "a social dilemma, in which everyone has a temptation to free ride instead of adopting the behavior that would lead to the best global outcome." According to them, "a serious consideration of algorithmic morality has never been more urgent." In what follows, however, we challenge both claims by plumbing the motivations for and upshot of the study. We are skeptical that algorithmic morality is a prerequisite to the rollout of AVs, and that the social dilemma is as intractable as Bonnefon *et al.* imply. Part 1 of the position paper is concerned with the importance of moral programming in AVs. Part 2 calls into doubt the alleged difficulty of en-

couraging people to buy utilitarian AVs—specifically, whether Bonnefon *et al.*'s results are an artifact of unrealistic experimental conditions.

Part 1: Putative Advantages of Moral Decision Rules

In their introduction, Bonnefon *et al.* claim that notwithstanding the low probability of collision scenarios,

"AV programming must still include decision rules about what to do in such hypothetical situations of unavoidable harm. Thus, these types of decisions need be made well before AVs become a global commodity. Distributing harm is a decision that is universally considered to fall within the moral domain. Accordingly, the algorithms that control AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm."

We want to argue that Bonnefon *et al.* close the book on the importance of moral programming too soon. It appears they are calling for AVs to be *explicit* ethical agents before entering the market. In other words, they are demanding that AV algorithms "represent ethical categories and perform analysis in the sense that a computer can represent and analyze inventory or tax information."[2] However, even supposing there is a consensus about which ethical categories to use, this is no simple task. While programmers work to embed AVs with moral principles, numerous social benefits—increasing traffic efficiency, decreasing pollution, and decreasing traffic accidents by 90%—are delayed.[1] It is worth asking then what are the advantages of designing AVs with moral algorithms, as opposed to algorithms yielding moral outcomes.

Can an AV still act in the hypothetical scenarios without an explicit *moral* decision rule to guide its actions? It would seem so. What distinguishes moral decision rules from non-moral decision rules is not just the 'calculation', but importantly the former's sensitivity to the putative moral features of a situation, such as the number of projected casualties. Certainly, an AV could still act in the absence of a moral decision rule; it simply wouldn't take moral features of the situation as input in its decision-making. It wouldn't act for moral reasons.

Perhaps, then, the more meaningful question is whether an AV can act *ethically* in a hypothetical scenario without an explicit moral decision rule. Here, it is important to distinguish implicit and explicit ethical agents (Moor 2006). One way to design a moral agent that does not contain moral decision rules is to create "software that implicitly supports ethical behavior, rather than by writing code containing explicit ethical maxims. The machine acts ethically because its internal functions implicitly promote ethical behavior—or at least avoid unethical behavior." So, AVs may still act ethically if algorithms embed *implicitly* moral decision principles, such as (crudely) 'always swerve away from human beings and aim for impact surfaces causing the least damage'. By also ensuring that AVs meet various structural criteria, such as effective crash zones, passenger restraints, and 'vision' systems for detecting pedestrians, manufacturers can create agents that minimize harm during car collisions and are therefore implicitly ethical, without designing them to have explicitly moral algorithms, i.e. algorithms embedding moral *principles* (e.g. 'maximize utility').

Here we arrive at a practical question. Perhaps Bonnefon *et al.* mean to say that AVs must include moral decision rules before entering the market because moral decision rules can protect manufacturers against charges of tort liability in situations of unavoidable harm—against charges that manufacturers have by an act or omission given rise to injuries amounting to civil wrongs in court. If so, we do not think this is a good reason for the urgency to design explicitly moral AVs. Including the moral decision rules must offer some legal advantage that would be forfeit without them. However, we argue that manufacturers would be liable for injuries caused in the hypothetical scenarios of Bonnefon *et al.*'s study *anyway*, regardless of whether the AVs had exercised moral decision rules. Given the low-probability nature of the scenarios surveyed in the study, it is likely that many AV functionalities would have to fail *before* the scenarios could be realized, such as failing to detect pedestrians from a sufficiently large distance away, failing to accurately calculate a no-casualty driving path from this safe distance, failing to deploy an emergency brake if no path is possible, and so on. Manufacturers would be liable even if the moral decision rules were non-defective.

On what grounds might someone sue the manufacturer for a non-defective decision rule that nonetheless led to an undesirable outcome? On the grounds that it is not the decision rule, but the *design* of the rule that is defective. This takes us to Part 2 of the essay. To conclude Part 1, then, let us reconsider the original question: "why do AVs need moral programming?" Not to act; not to act ethically; not to decrease the vulnerability of AV manufacturers to litigation. It is incumbent on Bonnefon *et al.* to describe the relative advantages of explicit ethical AVs over implicit ethical AVs and explain why the reasons canvassed in Part 1 are not exhaustive.

Part 2: A Flaw in the Experimental Design

A second major claim of the study is that manufacturing ethical AVs faces the challenge of satisfying three apparently incompatible objectives: "being consistent, not causing public outrage, and not discouraging buyers" (Bonnefon, Shariff, and Rahwan 2016). Programming less likely to cause public outrage is utilitarian and free of government regulation. But, according to the authors, utilitarian

programming appeals less to buyers compared to programming that prioritizes passenger safety, which without regulation has the consequence of pushing manufacturers of utilitarian AVs out of the market. Our objection to this line of reasoning concerns the assumption that utilitarian programming would appeal less to buyers. An alternative explanation of Bonnefon *et al.*'s results, in which a majority disfavored purchasing utilitarian cars for themselves, is that survey participants were not provided with important contextual information about the benefits and risks of AVs, and so were unable to make realistically informed decisions.

Imagine, for instance, if people were surveyed on their willingness to vaccinate and were told that one person's vaccinating would save 20 people, but would kill the person at some later time due to side effects. (The situation is not exactly analogous, since dying from side effects would not result in saving 20 people.) It would be hardly surprising if surveys found that a majority of participants were unwilling to vaccinate given the scenarios presented. This is because the perceived risk of vaccinating would be fundamentally altered. What would happen if participants were *also* told that the risk of dying from side effects is extremely low, that vaccinations protect individuals from contracting fatal diseases, and that vaccinating confers a social benefit by protecting the unimmunized and physically vulnerable? It seems unreasonable to expect that few participants would change their responses. After all, a major reason why many people vaccinate is that vaccines are safe and beneficial for personal and public health. A major flaw in the experimental design of Bonnefon *et al.*'s study is, thus, its failure to remind survey participants that the hypothetical scenarios are (we would argue extremely) low-probability, and that utilitarian AVs are likely to predominantly protect passengers and pedestrians alike. We anticipate that including these facts would change the distribution of responses and show that worries of a social dilemma are overblown.

None of this is to say that utilitarian AVs would be free from litigation. Undoubtedly, we can expect some victims of a non-defective explicit moral algorithm to sue manufacturers on the grounds that the moral algorithms have design defects, that the moral principles they embed are, in fact, immoral. The plaintiffs might endeavor to show that "the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design … and the omission of the alternative design renders the product not reasonably safe."[3] Such a charge of liability could pose a serious legal threat given that "a phenomenon called 'betrayal aversion' finds that people often have a strong emotional reaction against a safety innovation that actually causes harm" (Marchant and Lindor 2012). In the 90s, when General Motors (GM) attempted to defend the safety of its C/K pickup against the charge that GM's placement of the gas tank increased the risk of fatal fires after side impacts, comparative analyses showing that the placement of the tank yielded an equivalent or lower rate of fatalities from all accidents were received unfavorably by the jury. Nevertheless, if it is the case that, as Bonnefon *et al.* show, the majority find utilitarian AVs to be the morally preferable design choice, manufacturers might well be able to avoid unfavorable verdicts in such litigation cases.

Bonnefon *et al.* claim to have identified 'a social dilemma' given results from their studies, showing that a majority of survey participants took utilitarian AVs to be moral but disfavored purchasing the vehicles themselves. However, survey participants may have disfavored purchase because they lacked adequate contextual information about the social benefits and low risks of utilitarian AVs. Manufacturers, too, may have less reason to fear liability if utilitarian AVs truly have the backing of public moral opinion. That said, we hope to have shown in Part 1 that any advantages separating explicit ethical AVs from implicit ethical AVs are unclear. In the immediate future, then, it may be of best interest to the public good if we set aside questions of the comparative benefits of utilitarian and non-utilitarian AVs, and focus on releasing and distributing, with the greatest possible celerity, AVs whose social benefits are undisputed.

Implications for Cognitive Assistance and Government Regulatory Roles

The growing variety and number of systems for Cognitive Assistance are expected to yield numerous social benefits, such as increasing efficiency, decreasing environmental impacts, and decreasing accidents. While some have the view that manufacturers and regulators face a major design challenge of balancing competing public preferences. They advocate a moral preference for "utilitarian" algorithms; a consumer preference for vehicles that prioritize passenger safety; and a policy preference for minimum government regulation of vehicle algorithm design.

Our analysis for AV's above calls into question the importance of explicitly moral algorithms and the seriousness of the challenges identified. We conclude that the 'social dilemma' is probably overstated. Given that attempts to resolve the 'social dilemma' are likely to delay the rollout of socially beneficial AV's, we implore the need for further research and encourage manufacturers and regulators to commercialize AV's as soon as possible. We encourage the application of our analysis of AV's to additional areas of Cognitive Assistance research and product development.

References

Arkin, R.C. 2013. Lethal Autonomous Systems and the Plight of the Non-combatant. *AISB Quarterly*, 137: 4-12.

Balaban, C. 2014. Impacts of Autonomous Systems. In the Technical Report on the AAAI Fall Symposiun Series, November 13–15, at the Westin Arlington Gateway in Arlington, VA.

Bonnefon, J.; Shariff, A.; and Rahwan, I. 2016. The Social Dilemma of Autonomous Vehicles. *Science*, 352(6293): 1573-1576.

Nahamoo, D. 2014. Cognitive Computing Journey In PPAA '14 Proceedings of the first workshop on Parallel programming for analytics applications. 63-64.

Marchant, G. E.; and Lindor, R. A. 2012. The Coming Collision Between Autonomous Vehicles and the Liability System. *Santa Clara Law Review*, 52(4): 1321-1340.

Moor, J. H. 2006, The Nature, Importance, and Difficulty of Machine Ethics. *Machine Ethics*, 21(4): 13-20.