F1TENTH Autonomous Racing **(Due Date:)**

Lab 9: Robot Ethics

Instructor: INSTRUCTOR *Name:* STUDENT NAME, *StudentID:* ID

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission. All sources of material must be cited. The University Academic Code of Conduct will be strictly enforced.

1 Learning Outcomes

This assignment asks us to discuss some of the ethical considerations inherent in the design of autonomous systems generally, not just self-driving cars. It is divided into three parts, which ask the following:

1. Who or what is morally responsible for the decisions made by an autonomous system? (Part 3)
2. How can responsibility be programmed? (Part 4)
3. How do context and function affect the question of designing ethical behavior? (Part 5)

2 Overview

As engineers, we design systems to have a certain function: move parts in a warehouse, interpret and implement a surgeon's hand movements, or drive passengers around. Aside from functionality, we also design safety, security and reliability into our systems. With autonomous systems, we will be asked to design *ethics* into our systems, since they will share the world with us and make decisions autonomously, and some of these decisions are bound to involve ethical questions. How can we design and program an autonomous system that is ethical? Indeed, what does it mean for an autonomous system to be ethical? Or even, just *responsible* for its actions? Are there certain autonomous systems that simply cannot be ethical? Moreover, when speaking about designing ethical systems, *whose ethics are we talking about*?

The engineers who will be called upon to build and program autonomous systems must have an awareness of these issues and a solid foundation upon which to base their thinking about them. One can safely say that the engineer's own ethical obligation and professional code require them to think carefully about these questions and present their employers and the general public with

a principled analysis of the ethical dilemmas facing the autonomous systems they design - and in some cases, choose not to design.

Unlike other assignments and labs, this assignment does not require programming. Rather, you will read a number of papers, write response essays, and we will discuss the papers and your responses in class. *It must be emphasized that this assignment is as much a part of your education in autonomous systems and self-driving cars as any other, and is equally graded.*

2.1 Assignment Logistics

Half the groups in class will do parts 1 and 2. The other half of the groups will do parts 2 and 3. Next week in class, we will have a moderated discussion of these readings. During lecture time 1, we will cover parts 1 and 2. During lecture time 2, we will cover parts 2 and 3.

These articles are not easy - they are scholarly articles with sometimes elaborate arguments, which will require close reading and re-reading. Give yourselves enough time to do them justice, reflect on them, and discuss them amongst yourselves, before writing your response essays.

You are asked to write response essays in each part. A response essay outlines your thoughts on the matter in a *cogent, well-argued, well-written manner*, including your disagreements, if any, with the points made by the authors of the above readings. A response essay is not a free flow of thoughts: it makes a sequence of points and argues why they are valid. Thus when reading your essay, I should be able to tell, early on, what your thesis is. If you feel like you can't form a thesis yet, I should still be able to tell, early on, what are the competing theses that you are weighing. The essay is always about the topic under discussion: it does not branch out into marginally related musings. It also admits gaps in the writer's knowledge.

The response essays must be uploaded by 12pm on the day before class. (E.g., if class is on Monday, essays must be uploaded by 12pm on Sunday.) Essays uploaded after this time will not be read or graded. This is to give your teammates, and myself, a chance to read the essays.

Your grade in this assignment will take into account your response essays and the in-class discussion.

3 Where does responsibility lie?

Since the early days when it became apparent that self-driving cars are a real possibility, engineers, regulators and jurists started asking who would be liable when the car injures or kills someone though a mistake of its own? For instance, take the case of the Uber incident in Arizona, where an SUV in Uber autonomous mode struck and killed 49-year-old Elaine Herzberg on Sunday March 18, 2018[1]. Who is *responsible* for her death? Possibilities include:

- The car's passenger, though they were not in physical control of the car at the time
- The car's autonomy manufacturer (in this case, Uber[2])
- The designer of the computer vision algorithm that failed to identify the woman with sufficient confidence

[1] See the Guardian article here https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

[2] We are assuming the mechanical vehicle itself, manufactured by Volvo, is not at fault: the brakes didn't fail, the tires didn't slip, etc.

- The company that provided the dataset on which the computer vision algorithm (a deep neural net) was trained
- The city and state regulators who allowed Uber to test its cars on public roads at such an early stage

Note that the entity or entities responsible are not necessarily the ones that get sued in court, since the choice of who to sue depends on many other factors, including whether they're capable of paying damages, whether they're affected by negative publicity, etc. In particular, *moral responsibility*, which is what we're concerned with here, is different from *legal responsibility*, although a determination of moral responsibility is likely to affect the determination of legal responsibility.

We will explore this issue in our class, but in a different context, which brings out the urgency and salient traits of this issue in sharp relief: namely, in the context of autonomous weapons, that make the kill decision and implement it independently of humans. This is not because I think autonomous weapons are fine and dandy - I don't. Rather, it is because thinking about autonomous weapons makes it very clear what is at stake in terms of responsibility assignment. Who is morally responsible when an autonomous robot takes a decision to kill a human being?

3.1 Assignment

Your assignment for this week is individual. Every team member will do the following independently:

1. Read the following two essays, which you can find on the course website:
   - "Killer Robots" by Robert Sparrow, in *Journal of Applied Philosophy*, Vol. 24, No. 1, 2007
   - "Responsibility for Military Robots" by Gert-Jan Lokhorst and Jeroen van den Hoven, in *Robot Ethics*, Lin, Abney and Bekey (Eds), MIT Press.
2. Write a response essay, 2-pages long, single-spaced.
3. Read the response essay written by one of your teammates. Specifically, order yourselves alphabetically by your last name, and each person reads the response essay of the next person in the order (and the last person reads the first person's essay).

Discussion questions that we will raise in class for this part include:

1. Can we envision a scenario which, when viewed from the outside, leads us to say "this system is autonomous"? E.g., if the system is only told "Find enemy infrastructure and destroy it", vs. having a human guide the weapon to within 1km of a pre-specified target.
2. Recall the Commanding Officer sub-section of the "Killer Robots" paper (pages 9-10). Who is the equivalent of the commanding officer in the case of autonomous car? I.e., who 'orders' the car onto public streets? What does this imply for responsibility assignment?
3. Once we hold someone responsible for an action, do we then always follow that by reward or punishment? What are other purposes of assigning responsibility?

4 Responsibility Sensitive Safety

This part also concerns the question of responsibility as it relates to safety. We will consider one widely circulated proposal for defining responsibility for an accident involving an AV. The proposal was made in the following white paper:

- "On a Formal Model of Safe and Scalable Self-driving Cars" by Shalev-Shwartz et al.

The authors of this paper refer to this concept of responsibility as Responsibility-Sensitive Safety (RSS). The authors then propose that as long as an AV is programmed to never be responsible for an accident under the RSS definition of responsibility, then that is good enough. No statistical results need to be collected, no extensive testing needs to be done: just make sure, mathematically, that your control algorithms are never responsible for an accident according to RSS.

4.1 Assignment

Your assignment for this part is individual. Every team member will do the following independently:

1. Read the above paper, which you can find on the course website. You can skip the appendices.
2. Write a response essay, 2- to 3-pages long, single-spaced. Your response must argue whether Responsibility-Sensitive Safety (RSS), as defined in this paper, is a sufficient concept for the development of AVs. For instance, is RSS a sufficient specification for designing an AV's controllers? Could an AV satisfy RSS and yet lead to undesirable situations? If yes, describe them as concretely as possible. Is the notion of responsibility used in the paper sufficient? Does the driving policy proposed in the paper leave room for errors? If yes, how? Do you agree with the sensing model used in the paper?
3. Read the response essay written by one of your teammates. Specifically, order yourselves alphabetically by your last name, and each person reads the response essay of the next person in the order (and the last person reads the first person's essay).

The discussion in class will center around the questions that you address in your essay, and which are listed above in 2.

5 The context for designing moral decision-making algorithms

Another issue that is concerning jurists and autonomous cars' designers is that of *moral decision-making*. Assume that a child jumps in front of a self-driving car, such that only two courses of action are possible: either hit the brakes hard but still strike and likely kill the child, or swerve and hit the brakes but likely hit a wall and kill (or seriously injure) the passenger. What is the morally right decision? Of course, this decision is not unique to autonomous cars: it is one faced by human drivers as well, and early proposals seek to program certain rules of behavior that are supposed to be ethical. (As you might expect, what counts as ethical is subject to great philosophical debates, and one should not be lured by the simplicity of certain computational proposals).

In this part of the assignment, we explore the more fundamental question of whether certain autonomous systems, *by the very nature of the context in which they operate, or the nature of what they do, cannot be ethical*. For example, consider a cab company that offers rides to customers in

sedans (mid-size 4-seat cars). Their cars are autonomous. Seeking to project an image of luxury of itself, the company policy imposes that only one person can be in the cab. Now these cabs might have some ethics programmed into them concerning accidents. However, because of the one-person-per-car policy, the total number of cars on the road has now *increased*, making the pollution and traffic jams problems worse than before. It is then very unconvincing to argue that the car is ethical, given that its overall effect on the environment is actually harmful, regardless of how it is programmed. When considering whether a system is ethical, *it is important to consider the context it operates in, and the task it is doing, not just how it is doing it.*

Here too, we explore this question in the more stark context of autonomous weapons systems: what does it mean to claim that a robot has been programmed with ethical rules if the robot's task is to wage war?

Much of the thinking on creating 'ethical' autonomous killer robots starts from the theory of *Just War* ('just' as in 'justice', not as in 'just do it'). See here for a synopsis of just war principles: https://www.mtholyoke.edu/~jasingle/justwar.html.

5.1 Assignment

Your assignment for this part is individual. Every team member will do the following independently:

1. Read the following essay, which you can find on the course website:
   - "Killing Made Easy: From Joysticks to Politics" by Noel Sharkey, in *Robot Ethics*, Lin, Abney and Bekey (Eds), MIT Press.
2. Write a response essay, 2-pages long, single-spaced.
3. Read the response essay written by one of your teammates. Specifically, order yourselves alphabetically by your last name, and each person reads the response essay of the next person in the order (and the last person reads the first person's essay).

Discussion questions that we will raise in class for this part include:

1. Some, like Jeanette Wing, argue that it is in fact morally imperative to develop autonomous weapons systems, since they will reduce the chance of human error, whether in detecting, identifying, or targeting. In a nutshell, the argument is that if the autonomous weapon can kill just the targeted person, while a human wielding the weapon might kill the targeted person and three others, then it is our moral obligation to develop the autonomous weapon.

   What assumptions underlie this reasoning? What sequence of conclusions do we have to accept before we are at a point where the use of a weapon is inevitable? Are they valid assumptions and conclusions?

2. Some argue that the existence (and myth) of so-called smart bombs *increases* the likelihood of war because of the illusion that it is now 'cleaner' and better controlled. Analogously, the existence of safe autonomous cars might lead to *more* cars on the road since car transportation might come to be seen as safer and more reliable than, say, taking the bus or train. Discuss both of these points, for autonomous weapons and autonomous cars, and what unintended consequences might arise.

3. Talal Asad, in "Thinking about terrorism and just war" (*Cambridge Review of International Affairs*, Volume 23, Number 1, March 2010), argues that just war theory is very hard, if not impossible, to apply in today's wars, and indeed that the category of 'war', as usually understood, does not describe today's armed conflicts well: these conflicts have no discernible end or beginning, depend on a blurring of the boundary between combatant and non-combatant, etc. Thus one must ask: what is the meaning of programming the principles of just war if just war theory itself is inapplicable? How are some gaps and inconsistencies in our moral theories amplified when we apply autonomous systems? E.g., if just war theory depends on a distinction between combatant and non-combatant, and this distinction today is made by the commanders and the field soldiers, what happens when we field a robot that is inherently unable to make that distinction?

6 Further Readings

On responsibility in autonomous systems, you can start with the references under the section on Artificial Systems at: https://en.wikipedia.org/wiki/Moral_responsibility. You might even consider contributing to the page by summarizing your readings!

For an introduction to just war theory: https://www.iep.utm.edu/justwar/.

Anderson and Anderson explain how they programmed Nao, the Aldebaran robot, with ethical principles in the following article: Anderson and Anderson, "Robot be good", in *Scientific American, **303**, p. 72-77*, 2010

Recognizing autonomy in a robot requires us to first understand it in humans, which brings up the question of whether we humans have free will at all. For an argument that humans do not possess free will, see Sam Harris's short book, "Free will", Free Press, 2012. While Harris sometimes displays a shocking misunderstanding of politics and resistance, here he is at his finest.

For a defense by a philosopher of the idea that machines could be morally responsible for their actions, see D. Dennett, "When HAL Kills, Who's to Blame? Computer Ethics", in *D. G. Stork (ed.) HAL's Legacy: 2001's Computer as Dream and Reality (Cambridge, MA: MIT Press, 1997), pp. 351-365.*

7 Deliverables and Submission

Submit the following as `studentname_lab9.zip` (replace studentname with your name)

1. Responsibility sensitive essay titled `studentname_rss.pdf`
2. Moral decision-making essay titled `studentma_mdm.pdf`