F1TENTH Autonomous Racing

Robot Ethics - In Class Discussion Questions

Instructor: INSTRUCTOR

This lab and all related course material on F1TENTH Autonomous Racing has been developed by the Safe Autonomous Systems Lab at the University of Pennsylvania (Dr. Rahul Mangharam). It is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. You may download, use, and modify the material, but must give attribution appropriately. Best practices can be found here.

1 Overview

These are suggested arcs for conducting in-class discussion that accompany the 'Moral Decision Making' lab.

2 Where does responsibility lie?

For the reading

- "Killer Robots" by Robert Sparrow, in *Journal of Applied Philosophy*, Vol. 24, No. 1, 2007

The suggested arc of questions is

1. What is the thesis developed in Sparrow?
   A: From the abstract: "I argue that in fact none of these [loci of responsibility] are ultimately satisfactory. Yet it is a necessary condition for fighting a just war, under the principle of jus in bellum, that someone can be justly held responsible for deaths that occur in the course of the war. As this condition cannot be met in relation to deaths caused by an autonomous weapon system it would therefore be unethical to deploy such systems in warfare."

2. What is the subject of the first section of Sparrow?
   A: an overview of current and near-future military systems that have some autonomous capabilities.

3. At what point would you consider a system to be 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, vs. guiding it to the final target and impact point.

4. Sparrow says that "Military interest in UCAVs stems from two main sources. Firstly, uninhabited systems offer the prospect of achieving military objectives without risking the politically unacceptable cost of friendly casualties. Secondly, they are expected to be substantially cheaper than the piloted systems they are intended to replace". What is the implicit value system that allows this calculation of political cost? Whose lives must be protected according to this calculation?
   A: the lives of 'our' soldiers are worth more than N lives of the population being attacked. Note this isn't about our soldiers vs. theirs - it is about our soldiers vs their population, since today's wars are not restricted to clearly demarcated battlefields. (In fact, as early as WWII, cities were bing bombed in Europe, and way before in colonized states).

5. Is there a similar dynamic in self-driving cars, where one group's interests are advanced over another's?
   A: e.g. the interests of upper middle-class people who can afford these cars and a house in the suburbs over those of poorer people who live in the city: the city tax breaks go to companies developing AVs rather than to public transport, or to companies developing self-driving technology for public transport.

6. Read the opening example of p.5. Hypotheticals are a very useful way of analyzing situations and distinguishing the important factors at play.

7. Recap why it's important to be able to assign responsibility.
Respect for the enemy; being able to apply the principles of jus in bello.

8. Is there a parallel obligation in AVs to be able to assign responsibility for the willful killing of a pedestrian, say?
A: yes: respect for others' inalienable dignity, and being able to apply laws.

9. "There is an obvious tension involved in holding that there are good military reasons for developing autonomous weapon systems but then not allowing them to fully exercise their 'autonomy'." Is there a similar tension in Tesla's so-called Auto Pilot between reasons for using autonomy, and promises to limit its use?
A: Tesla's so-called Auto Pilot can presumably drive on the highway, but Tesla asks the driver to remain engaged at all times, while making it harder for the user to engage, by reducing their cognitive workload.

10. Sparrow lists three candidates for the locus of responsibility that he consider exhaustive: the programmer, the commanding officer, and the machine itself. Are there others? Can the programmer be taken as meaning 'the company', and the commanding officer as meaning 'the government'?

11. Recap the 2 arguments against holding programmers responsible.

12. 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?

13. Recall the argument against holding the machine responsible. This is a *reducio ad absurdum*: take the reasoning to its logical conclusions and reach a contradiction. If the machine can be said to genuinely suffer, then we should be concerned about its pain in war, and should be as reluctant to send it into war as a human. This defeats the purpose behind building it!

14. Similarly, a truly suffering machine would also feel empathy and would not be a cold killing machine.

15. 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?

16. What are possible loci of responsibility for AVs?

For the reading

- "Responsibility for Military Robots" by Gert-Jan Lokhorst and Jeroen van den Hoven, in *Robot Ethics*, Lin, Abney and Bekey (Eds), MIT Press.

The suggested arc of questions is

1. Does Sparrow really assume that all AWS (Autonomous Weapons Systems) are 'killer robots', as Lokhorst and van den Hoven (L&H) claim in this essay?
A: No. Sparrow focuses on the case of a war crime to remove all ambiguity that a crime has been committed and ground the discussion in this fact. He does not claim that all AWS have killing as a purpose. This is an example of a 'strawman argument'.

2. The authors, in Thesis 1, speak of robots that 'save lives'. Their goal is to main and injure - in what sense are they saving lives? What if the more efficient robots maim more people than humans kill or injure? Moral: always examine the assumptions and the context and modify the parameters of a thesis.

3. Does the threat of punishment really serve no purpose at all? E.g. what is the robot (just like a human) is capable of faking intent - e.g., lying. Then it can reason that if it is caught doing the wrong thing, it can pretend to have erred, we fix it, and then repeat. But if there is a threat to definitely punish a wrongdoing, then that might serve as a deterrent.

4. If the ability to punish (or be punished) is not necessary, then what distinguishes an AWS and a long-range weapon? If the long-range weapon malfunctions, you also fix its design, etc.

5. Is the prospect of debugging an autonomous AI realistic, even 100 years from now?
A: obviously, we should be cautious in answering this. But consider that today it takes months to debug a silicon bug in a microprocessor. And an AI will have 10s of microprocessors, underneath everything else. How long will it take to debug that?

6. Sherry (a student) made the following good point: that the people responsible should be the people who are most capable of ensuring system safety, since this will motivate them to do their job, as it were. Holding the owners responsible does pose a challenge: the owner buys the car because it is autonomous and they can sleep, say, in traffic. If the law requires them to be alert at all times, doesnt this defeat the purpose of buying the car?
A: Of course, this is not an ethical question, it is a commercial question that need not concern us here. Ultimately, if whats ethical and whats commercial conflict, we should take the ethical position.

3 Responsibility Sensitive Safety

The reading is

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

The suggested arc of questions is

1. This might've seemed like too much to read. But if you join an AV company, it is likely they will ask you to read it, evaluate it, and report back. So this is training for doing that.

2. What are the 2 claimed contributions of the paper?
A: from the intro, a formalization of the "common sense of human judgment with regard to the notion of who is responsible for causing an accident", and a "semantic language for units, measurements, and action space and specifications as to how they are incorporated into Planning, Sensing and Actuation of the AVs".

3. Do you agree that typically only "one of the agents is responsible for an accident"? Do they present evidence for that?
A: They don't present evidence - we have to examine accident databases. Also, this depends on the horizon we examine before the accident moment. It might be that at the last minute one car is to blame. But what about the sequence of events that led to that moment?

4. Go over the probability calculation in Section 2.2 and Corollary 1.

5. There is a slight error in the discussion about validating a simulator. What is it?
A: a policy for an AV need not be simulated - we can load the code of the AV into the simulator and run that. The simulator simulates ***the environment fo the AV including its sensory inputs****. When the policy is changed, we need to update the simulator to simulate reactions to the new policy. There are no formal calculations here, and I don't see in what sense validating the simulator is as hard as validating the policy.*

6. What is Soundness? Usefulness?

7. Are Usefulness and Soundness equally important?

8. What about Completeness: does the model find all cases where the car is to blame?

9. Do you agree with the re-definition of safe distance for cars traveling in opposite directions?

10. Note: they don't give concrete values for the numbers. Be suspicious of this.

11. What is the dynamical model used to derive the equations for safe distances?

12. Note: don't hide propositions (that need proof) inside definitions, like they do in Def. 3.

13. Could car 1 be responsible for an accident with car 2 according to Def.5, but only because it didn't want to be responsible for another accident with car 3. Let's make this concrete. Should car 3 be responsible for both accidents?

14. How exactly are they measuring lateral and longitudinal distances? By Pythagoras? Def.9 depends crucially on this. Can both distances really be unsafe?

15. Def. 9 case 2 bullet 1: t is any dangerous time it seems. Therefore, by definition of blame time, there is *always* a dangerous time t in $[t_b, t_b + \rho)$ since *every time in* $[t_b, t]$ *is also dangerous.* So what are they really talking about here?

16. Itemization following Def. 22 (Sensing system types of errors): this assumes implicitly that the perception system's only task is to detect objects. Discuss.
    A: perception also serves to evaluate the state of the road (slippery, rough, provides good grip, etc), traffic signs, road physical constants (width, elevation gains, etc). These can have continuous measurement errors.

17. They say they perform "offline validation [of the sensing system] after every major update of the sensing system". What about a lens change? is that a 'major update'?
    A: yes. Nvidia had to re-train their neural nets.

18. Is RSS (Def. 10) a sufficient specification for designing an AV's controllers?

19. Could an AV satisfy RSS and yet lead to undesirable situations? If yes, describe them as concretely as possible.

20. Is the notion of responsibility used in the paper sufficient?

21. Do you agree with the sensing model used in the paper?

22. Is this a proposal to define responsibility, or to escape legal liability?

4 The context for designing moral decision-making algorithms

The readings are

- "Killing Made Easy: From Joysticks to Politics" by Noel Sharkey, in *Robot Ethics*, Lin, Abney and Bekey (Eds), MIT Press.
- A synopsis of just war principles: https://www.mtholyoke.edu/~jasingle/justwar.html.

Suggested discussion arc:

1. What are the principles of just war?

2. What are the points made in Sharkey's essay?
   A: Sharkey argues
   1) that killing from a distance is appealing to the generals because it is cheaper and to the politicians because it is more visually appealing (no body bag count)
   2) that AI cannot discriminate between combatant and civilian because we don't have a definition of civilianness, and even if we did the AI can't get access to all the information needed to make a computation, and this is an essential distinction in the laws of war
   3) that Arkin's proposal of an ethical code is bogus because it can't be implemented (it requires access to information and ability to do computations that will never be available)
   4) that a robot that doesn't seek revenge or feel hate also cannot feel empathy and kindness
   5) that the information the military gets is tainted by errors and willful deceit by informants

3. On the first point: shouldn't war be 'visually' horrible, shouldn't we know exactly what is being done in our name, both to our co-citizens and other human beings, if we believe at all that all humans are equal? 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.

4. On the second and third points: What do you think of the argument that says that distinguishing combatants and civilians is a complex process that cannot be done by machine? E.g. signature strikes: “targeting people whose behavior is assessed to be similar enough to those of terrorists to mark them for death with criteria that can be as vague as killing “military-aged males” in regions where terrorists operate”.[1]

5. The burden of the proof is on the person stating the affirmative.

6. On the fifth point: an AWS can be a weapon of mass destruction. What is the consequence of tainted or deceitful information?

7. 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?

8. What about this scenario: the government doesn’t program any restrictions on its robot army. After a massacre, they upload remotely a software that does have these constraints, and claims that its robots could not have done this because hey, constraints!

[1] https://www.theguardian.com/us-news/2016/jul/01/obama-continue-signature-strikes-drones-civilian-deaths