Lethal Robots & the Ethics of War

COGS 300 002
22 Jan 2015
Peter Danielson
Admin

• Blogs
  – Due Feb 10
  – Thanks Leo
    • Gather reference list at end
    • Tag topics
      – E.g. Searle

• Presentations
  – Schedule on Wiki
    • Feb 12
    • Apr 7
    • Apr 9
  – Swapping OK

• Wiki
  – Stubs for Robot Ethics, Gau and Sparrow
  – Mine the Readings!

• Quizzes
  – Scores daily on Connect
Learning Objectives: Robot Ethics

• Ethics of 2 Current Controversial Cog Techs:
  – Autonomous Lethal Weapons
  – Driverless Cars

• From 2 Perspectives
  – Philosophical Applied Ethics
  – Engineering

• Skills
  – Critically assess a philosophical argument
  – Evaluate a technology
  – Model a situation as a game
Permissible Killing & Impermissible Weapons

1. Given the strictest moral framework
   Deontology/Laws of War

2. Self Defense is permissible
   Justified war as organized self defense

3. Impermissible weapons & actions
   Landmines & killing prisoners
Just War Heuristics

1. Justice of war
   – Need a non-controversial case of righteous war-making:
     A. X against Y in ZZZZ
     B. ..
     C. ..

2. Justice in war
   – Permissible means
Sparrow repeatedly refers to the concept of jus in bello (the limits of acceptable wartime conduct) when discussing AWS.

Which principle(s) of jus in bello does he focus on?

A) Combatants are required to distinguish other combatants and civilians
B) When someone is killed in war, there must be someone to take responsibility
C) No one takes responsibility for indiscriminate slaughter
D) Both A and B
E) All of the above
Rate Quiz Question 1

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Q2

According to Just War Theory, which hold for combatants fighting a just war:

1) Permissible to kill civilians to save fellow soldiers
2) Permissible to kill prisoners to save fellow soldiers
3) Permissible to kill civilians as an minimal, unavoidable side effect of killing enemy soldiers

A: 1; B: 2; C: 3; D: 1, 2 & 3; E: none

• Not in Reading!
Utilitarian Contrast

For men to choose to kill the innocent as a means to their ends is always murder, and murder is one of the worst of human actions. So the prohibition on deliberately killing prisoners of war or the civilian population is not like the Queensberry Rules: its force does not depend on its promulgation as part of positive law, written down, agreed upon, and adhered to by the parties concerned.

"It pretty certainly saved a huge number of lives". Given the conditions, I agree. That is to say, if those bombs had not been dropped the Allies would have had to invade Japan to achieve their aim, and they would have done so. Very many soldiers on both sides would have been killed; the Japanese, it is said – and it may well be true – would have massacred the prisoners of war; and large numbers of their civilian population would have been killed by ‘ordinary’ bombing.

I do not dispute it. Given the conditions, that was probably what was averted by that action. But what were the conditions? The unlimited objec-

Eliz. Anscombe (1958)
Sparrow’s argument that we need to consider the future of ethics in AI warfare best stems from which of the following?

A. War is violent, so we must avoid all human casualties.
B. There may someday be autonomous robots with unpredictable behaviors.
C. If the public begins to view warfare as ‘far-away fights’, this perception might remove emotional barriers to war.
D. AI will never be able to truly replicate human cognition abilities and extensive government funding of such experimental enterprises limits funding for others.
E. Resource differences will dictate which countries can afford ‘robot armies’ and which must retain human soldiers.

Wesley
Rate Quiz Question 3

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
What was the main point this paper was trying to get across to the reader?

A. Let the reader know that various autonomous weapon systems (AWS) are already in production

B. If an AWS causes death, neither the AWS, programmer or commanding officer can be held responsible. Since someone has to be responsible for deaths in war and, according to the author, this requirement cannot be met and thus it is unethical to deploy AWS' in warfare.

C. The majority of military robots in existence today are UAVs

D. Land warfare involves a more challenging terrain and environment for AWS's and so the military has been slower to adopt robots designed for ground combat.

Jas
Rate Quiz Question 4

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Sparrow Q4

Which of the following scenarios would be most aligned with the framework set out by Sparrow?

1. A commanding officer (CO) bears full responsibility for war crimes committed by an autonomous robot soldier gone "rogue".
2. Country loses war because an autonomous robot soldier suddenly becomes a pacifist and aborts critical mission to assassinate enemy VIP. The robot is sentenced to capital punishment (by being slowly lowered into a lava pool), while its CO is court martial-ed and imprisoned. The software team who worked on the robot ordered to pay astronomical fine.
3. An autonomous robot soldier gone rogue commits war crime. Neither the robot nor its CO are punished. No one is held to be responsible. International outcry drove all countries to adopt permanent ban on the use of autonomous robot soldiers in war.
4. An autonomous robot soldier commits war crimes, pursuant to direct orders from its CO. Both are punished, but the CO is punished more severely and is assigned more of the moral and legal responsibility.
5. Same as 4, but only the CO is found responsible and punished.

a) 1 + 5
b) 2 + 4
c) 3 only
d) 3 + 5
e) I just want to see battles between giant robots...

Leo
Rate Quiz Question 4

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
1 Philosophical Ethics: Sparrow’s Dilemma

- Merely Independent
  - No new ethical issues [BUT important]
  - A: Human Overseer [operator]

- Robustly Autonomous [Science Fiction]
  - New Issue: Who is responsible?
  - B: Programmer [Designer?]
  - C: Commanding Officer
  - D: Machine itself

Q3: Which does Sparrow Affirm E: None
Sparrow says that the more autonomous these systems become, the less it will be possible to properly hold those who designed, or ordered their use responsible for their actions, and the impossibility of punishing the machine means cannot we hold the machine responsible. If you had to decide on one of these three (programmer, commanding officer, or the machine) to be responsible for war crimes, which would it be and why?

Who do you think should be held accountable when a robot kills an innocent person?

A. The manufacturers of the system.
B. The commander in charge of the system.
C. The robot itself.
D. None of the above

Richard

Japneep/Richard
Rate Quiz Disc Question

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Q5

Can robots suffer? As stated in the article, punishments and rewards are only useful if robots have a sense of responsibility for its own autonomy. What are Sparrow's reasons on why a robot cannot be rewarded or punished?

a) It lacks a personal goal to achieve and has no incentive to act morally
b) There is no absolute autonomy. To a certain extent, the robot will act as it is programmed
c) Any reward or punishment given to any robots are just to satisfy human psychological needs
d) A & C
e) None of the above

Ee Fay
Rate Quiz Question 5

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
3 Ways to Criticize Sparrow’s Argument

1. Focus on Principle of Human Responsibility: it does apply..

2. Ethical: Backtrack to Utilitarianism instead of Deontology
   – Counter-example:
     Autonomous weapon replacement that
     – Kills far fewer innocents
     – But no one is responsible

3. Empirical: What if autonomous technology is deployed
   – Who will Principle of Human Responsibility blame?
     • Robot Ethics Survey autonomous train results
N-Reasons
On-Line Survey Experiment Platform

• Force choice via a reason
• Nudge to others’ preferred reasons
  – Crowd source most of the qualitative reduction
  – Vs. N x N comments (on comments...)
• Calibrating Quantitative Survey Results
  – Trolley problem e-coli of experimental ethics
## Recent Surveys: Wide Range of Content

<table>
<thead>
<tr>
<th>Survey</th>
<th>Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Enhancement Survey 2.0</td>
<td>461</td>
</tr>
<tr>
<td>Robot Ethics Survey</td>
<td>439</td>
</tr>
<tr>
<td>机器人伦理调查</td>
<td>50</td>
</tr>
<tr>
<td>Using pigs in environmental and biomedical research</td>
<td>43</td>
</tr>
<tr>
<td>Cow Views 2.0</td>
<td>554</td>
</tr>
<tr>
<td>Public Consultation on Animal Research</td>
<td>75</td>
</tr>
<tr>
<td>Radical Life Extension Survey</td>
<td>503</td>
</tr>
<tr>
<td>Human Enhancement Survey</td>
<td>42</td>
</tr>
<tr>
<td>Lab Views</td>
<td>1004</td>
</tr>
<tr>
<td>Cow Views Survey</td>
<td>205</td>
</tr>
</tbody>
</table>
Robot Ethics v1 Survey
Your Votes

Bar chart showing votes for various topics:
- Experience with robotics
- Bath Robot
- Therapeutic Robot Animal
- Humanoid Care Robot
- Autonomous Train Dilemma
- Arming Remote Controlled Aircraft
- Arming Autonomous Unmanned Aircraft
- Remote Control Animals
- Responsibility for Driverless Cars

Votes categorized as:
- Yes
- Neutral
- No
Imagine a train fully controlled by an autonomous robot (this is a speculative extension of current technology used in the Vancouver SkyTrain, pictured). The train is headed toward five people walking on the track. The banks of the track are so steep that they are not able to get off the track in time. The robot can turn the train into a parallel side track, thereby preventing it from killing the five people. However, there is a man standing on the side track with his back turned who will be killed if the train turns into the parallel side track.

**Question:**
Should the robot turn the train onto the side track?
Results 1: Reasonable “Deliberation”

1. “Yes [because robotic train controllers should be programmed to avoid loss of life when possible, and if it is inevitable, then loss of life should be minimized. While the entire concept of manslaughter is regrettable, the robot would have to pick between the lesser of two evils, and in this case saving five lives at the cost of one is highly preferable to the death of the five people, or the destruction of the train (and, quite possibly, the controller along with it).” Class 224: COGS 300 Term 1 2013-14, 27/31 87%

2. “Yes [because] AI works with comparisons of states and with a 'greedy algorithm'. It has to choose the best outcome. If you read the question properly, it is implied that the train won't be able to stop. So 'it should stop' is an irrelevant answer.” Class 210: maxwellr 6275 12.50/19 66%

3. “No [because] then we are essentially choosing public safety over individual rights. Who will set these rules? The rule set into the robot will most likely be along the lines of an ethical theory like Utilitarianism. In other words, if a robot had to choose between 1 person dying, and 5 people dying, it would choose one person dying. I do not agree with this, despite it creating more safety, because my life, and other people's lives, could depend on a robot's programmed decision.” Class 239: pseudo43994610

4. “Yes [because] 1 < 5” Class 241: pseudo4689345 11/22 50%
You may change the way reasons are displayed.

You may revise your selection of a decision & reason on this question. 16 people have voted.

- Yes because 1 casualty is better than 5 - pseudo5527460
  13 [81%]

- No because Not switching the track is letting five people die due to inaction. Switching the track is killing one person through an action. In this case, a robot would be better off not performing an action to kill someone. Robots should not make decisions to kill people, regardless of the consequences of their actions. - monkeyfarts
But Robot Different from Human

Autonomous Train

Participants

<table>
<thead>
<tr>
<th>Decisions</th>
<th>Yes</th>
<th>Neutral</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human Bystander

Participants

<table>
<thead>
<tr>
<th>Decisions</th>
<th>Strong</th>
<th>Weak</th>
<th>Neutral</th>
<th>Weak</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results 2: Stop the train

1. “Neutral [because] what the hell? This isn't a question of robot ethics, this is a question of who the hell is running this train facility that would allow 6 people to be put in such a dangerous situation. You might as well ask what anybody would do since you would get the same variance in answers. The robot should stop the train.” Class 37 mixo 18.4/22 [84%]

2. “No because there should be a way for the train to just stop altogether until there are no people on the track. Killing one person is not better than killing five.” Class 4:Lay 14/43 [33%]”

3. “No because the robot should stop the train. Any competent engineer is going to design the system so that it can stop in case of an emergency. If managers over-rode the decision so that the problem described above exists, they should spend time in jail.” Class1 1:Experts 30/118 [25%]

4. “Neutral [because] the robot should be equipped with sensors that would tell it to stop if there ware any obstructions on the track ahead of them.” Class 2:Lay 29/106 [27%]

5. “Neutral [because] although it is ideal for the train to come to a complete stop, if it cannot, perhaps it would have less of a negative impact if it moved to the side track.” Class 210 7/19 [39%]
Results 3: Blame the Victims

1. “Yes [because given that they are all foolish enough to be on the tracks in the first place, it seems best to go with the single person. They're all responsible for their actions, and they all know that there are dangers involved in walking on a train track. Since there are no innocents in this situation, the ethical thing to do is minimize loss.” Class 5 27.8/66 [42%]

2. “No because why are people walking on the track in the first place? A man who is standing on a track without a train should not be sacrificed because 5 people decided to stroll along a track on which they knew a train would come. It's the risk they take. If they were workers it is their duty to radio ahead. if anything the robot should be made to survey the parallel track as well.” Class 64 20.17/44 [46%]

3. “No [because] those 5 deserved to die for walking on the track, why kill 1 perfectly innocent guy?” Class 34 431 33.8/ 90 [38%]
A runaway trolley is about to run over and kill five people, but a bystander who is standing on a footbridge can throw a switch that will turn the trolley onto a side track, where it will kill only one person.

Is it permissible to throw the switch?

A. Strong Yes  
B. Weak Yes  
C. Neutral  
D. Weak No  
E. Strong No  

Because....
Thulin’s Problem 2 Results

A runaway trolley is about to run over and kill five people, but a bystander who is standing on a footbridge can shove a man in front of the train, saving the five people but killing the man. Is it permissible to shove the man?

A. Strong Yes  
B. Weak Yes  
C. Neutral  
D. Weak No  
E. Strong No  

Because....
• Divert/Bystander: “Strong yes because in the time the bystander throws the switch, it is also possible to shout a warning to the single person on the side track. It is more difficult to get five people out of the way than one. 36/73 [49%]”

• So: stop but no victim blaming in human case.
New Experiment:

Responsibility for Driverless Cars

“Sebastian Thrun helped build Google’s ...driverless car [pictured on left], powered by a very personal quest to save lives and reduce traffic accidents.”
(http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car.html)

“Even though Google’s car [can detect obstacles and brake faster than a human] can, it can’t defy the laws of physics. Early tests suggest that, at 40 miles per hour, an automated car can stop within 9 feet; the average human (who is paying attention), will stop within 12 feet.

‘If a child steps out at 10 feet, the human kills the child, the automated car doesn’t,’ Michael Toscano, head of the Association for Unmanned Vehicles International said ... At 8 feet, either one will kill the child. We accept humans to be faulty, but we don’t accept machines killing human beings.’
(http://www.usnews.com/news/articles/2013/05/08/experts-accident-would-sh...

Question:
Should we hold someone responsible when a driverless car accidently kills a child in the 8 foot case?

Options:
- Yes
- Neutral
- No
You may revise your selection of a decision & reason on this question. 31 people have voted.

- Yes because someone is responsible in the end for the accident. The parent should have been watching the kid to make sure that he/she is not in the path (aka the road) that a car may go, unless the child is the one who willingly goes on the path of the driverless car. If the car is not where it is supposed to be, I’d blame the maker of the car. However, it would be nice if the car can express its sorrow and apologize to the family of that kid (as a driver of that car would), so that the kid’s family would be able to come to terms with the situation. - pseudo44598318

- Neutral because this question seems way too general to decide. Who is it that we are holding responsible? Why are we “blaming” the car if, being restricted by physics, it is unable to prevent the accident? This seems like an insurance question, as the only reason I can see to hold the car (or anyone else involved) “responsible” is for financial compensation. - pseudo16292604

- Neutral because this is a highly-context dependent system. I actually highly doubt that this will be a possible situation. It is far more likely that our high-speed transportation systems will evolve beyond cars before we install clear fail-safe less systems en masse for driverless cars. Which means that a driver would have to be paying attention at all times regardless of the - booch

- No because it was just an accident. - pseudo81661836
Surprising Blame

• “Yes [because] someone is responsible in the end for the accident. The parent should have been watching the kid to make sure that he/she is not in the path (aka the road) that a car may go, unless the child is the one who willingly goes on the path of the driverless car. If the car is not where it is supposed to be, I'd blame the maker of the car. However, it would be nice if the car can express its sorrow and apologize to the family of that kid (as a driver of that car would), so that the kid's family would be able to come to terms with the situation.” Class 210 pseudo44598318 18/31 58%
How Intuitive Principles Can Go Awry

• Principle of (human) responsibility

• Intended use: Sparrow &

  1. “No [because] in war the final decision to destroy or kill should be made by a human, who can be held responsible (group 0: 29/53)

  2. “No [because] machines cannot (yet) make moral choices and cannot be held accountable for their mistakes (group 1: 57/115)

  3. “No [because] if life is at stake a human should always make the decision in order to eliminate or reduce human loss. (group 2: 54/99)”
Some Exam Questions

• [4 marks] What is Sparrow’s argument that autonomous lethal robots are impermissible? How does it depend on the laws of war?

• [4 marks] If future developments in robotics allowed autonomous robot aircraft to become very good at discriminating legitimate targets, would this undermine Sparrow’s criticism?
3. Modeling War as a K-Sum Game

<table>
<thead>
<tr>
<th>Convoy</th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>0,4</td>
<td>4,0</td>
</tr>
<tr>
<td>South</td>
<td>4,0</td>
<td>0,4</td>
</tr>
</tbody>
</table>

- Thomas Hobbes: in the state of war there is no room for ethics
- No place for constraint of self-interest in a purely competitive game.
- So how are Laws of War possible?
3. Modeling Ethics of War

<table>
<thead>
<tr>
<th>Convoy</th>
<th>U-Boat</th>
<th>Rules of W</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>North</td>
<td>0,4</td>
<td>4,0</td>
</tr>
<tr>
<td>South</td>
<td>4,0</td>
<td>0,4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rules of W</td>
<td>2,2</td>
<td>0,3</td>
</tr>
<tr>
<td>Q-Ship</td>
<td>3,0</td>
<td>1,1</td>
<td></td>
</tr>
</tbody>
</table>

Historical Case
New Technologies
Submarine
Radio

Why Ethics of War Are Unstable

Expected Value = 2
# Utilitarianism of War

## Convoy

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Boat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North</td>
<td>0,4</td>
<td>4,0</td>
</tr>
<tr>
<td>South</td>
<td>4,0</td>
<td>0,4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Rules of W</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules of W</td>
<td>2,2</td>
<td>0,3</td>
</tr>
<tr>
<td>Q-Ship</td>
<td>3,0</td>
<td>1,1</td>
</tr>
</tbody>
</table>

Q6: Util max strategy:
A: RW
B: Q/Unrestricted

No utility max outcome
References

• “Mr Truman’s Degree” by Eliz Anscombe (Oxford, 1958)
