Surprising judgments about robot drivers: Experiments on raising expectations and blaming humans

Peter Danielson

Centre for Applied Ethics, School of Population & Public Health, University of British Columbia, Vancouver, Canada, danielsn@exchange.ubc.ca

N-Reasons is an experimental Internet survey platform designed to enhance public participation in applied ethics and policy. N-Reasons encourages survey respondents to generate reasons to support their judgments, and groups to converge on a common set of reasons for and against various issues. In the Robot Ethics Survey, some of the reasons included surprising judgments about autonomous machines. Participants gave unexpected answers when presented with a version of the trolley problem with an autonomous train as the agent, revealing high expectations for the autonomous machine and shifting blame from the automated device to the humans in the scenario. Further experiments with a standard pair of human-only trolley problems refine these results. Responses reflect high expectations even when no autonomous machine is involved, but human bystanders are only blamed in the machine case. A third experiment explicitly aimed at responsibility for driverless cars confirms our findings about shifting blame in the case of autonomous machine agents. We conclude methodologically that both sets of results point to the power of an experimental survey-based approach to public participation in exploring surprising assumptions and judgments in applied ethics. However, these results also support using caution when interpreting survey results in ethics and demonstrate the importance of qualitative data to provide greater context for evaluating judgments revealed by surveys. On the ethics side, the result about shifting blame to humans interacting with autonomous machines suggests caution about the unintended consequences of intuitive principles requiring human responsibility.

Keywords: autonomous machines, trolley problem, robot ethics, responsibility, survey methods

Introduction

Theoretical Background

This article suggests that survey-based research holds great promise for ethics. In metaethics, survey and fMRI methods in moral psychology and experimental philosophy have supported new approaches to accounting for intuitive judgments in ethics (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, 2013; Mikhail, 2007; Hauser, 2006). In applied ethics, we have argued that surveys support a more democratic basis for ethical deliberation (Ahmad et al., 2006; Danielson, Ahmad, Bornik, Dowlatabadi, & Levy, 2007; Ahmad, Bailey, & Danielson, 2010). More modestly, our open-ended survey instrument allows exploratory access to moral judgments that might be ignored when we focus on debates between

theoretically structured alternatives. Indeed, this is precisely what happened in the case reported in this paper. Our long-running Robot Ethics Survey included the Autonomous Train Dilemma, a robot version of the basic trolley problem (see Figure 1). The scenario did not result in the expected distribution of judgments; it led instead to what we interpreted as protest "votes". Some respondents explicitly rejected the question as having nothing to do with robot ethics, others chose the Neutral response to avoid the dilemma, and others gave extreme reasons for their choices. In several papers, our team deemphasized this question (one of 9) in our analysis of robot ethics as seen through our survey (Danielson, 2011b; Moon, Danielson, & Van der Loos, 2012).

However, we should welcome surprising data, especially in ethics. In the present paper we argue that the unexpected participant judgments reveal important issues, methodologically about survey research and experimental scenarios in ethics, and ethically about introducing autonomous machines into our morally structured interactions.

Research Questions

The unanticipated responses to the Autonomous Train Dilemma led to the research questions that we explore in this paper. First, can we categorize the qualitative data provided by participants as interesting new reasons, or should we dismiss the Autonomous Train Dilemma as an unreliable stimulus? Second, given that the qualitative data reveals surprising new reasons, are these reasons unique to the robot agent case or are they also found in a second survey that explored the standard, human agent trolley problem? Third, having identified blaming human bystanders as a special problem for the robot agent case, we created a new question for the Robot Ethics Survey about an accidental death involving an automated vehicle. Would participants find humans responsible even in this new, more extreme case? We address these questions in Experiments 1 – 3, respectively. The larger research question is whether experimental survey research can contribute to our understanding of applied ethics.

Methodology used

The methodology we use to answer these questions is Internet-based survey research based on scenarios used in moral psychology and experimental philosophy. The first key innovation in our methods is to require participants to provide or select reasons for their answers, linking richer qualitative data to survey responses (Danielson, 2010). The second innovation is to use a constant survey framework for many groups and surveys, allowing us to compare responses of different groups and to new scenarios as we refine our research questions.

Methods

N-Reasons Survey Platform

The N-Reasons survey platform was designed to provide a bridge between clear, readily interpreted quantitative data and richer qualitative results. In particular, we can summarize and report groups' decisions in broad normative categories (e.g. Yes, Neutral, No – see Figure 2) but we can also drill down into the various reasons that participants contribute to support their decisions – see Table 2 (Danielson, 2011a; Danielson, 2013). By presenting participants with reasons combined with vote counts (see Appendix), our platform aims for the ideal normative procedural goal of reflective equilibrium, wherein each participant chooses their response in light of all other participants' choices and reasons. Moving from the ideal to the

feasible, we need to avoid the proliferation of qualitative data. First, participants can opt to concur with reasons contributed by other participants, rather than contributing new variations. By combining quantitative and qualitative data sources, we can examine the most popular reasons given by the 826 participants in the Robot Ethics Survey without the arduous and methodologically complex task of reading and classifying 826 comments for each question. For the Autonomous Train Dilemma, only 112 reasons attracted votes from participants other than their authors; this is an order of magnitude reduction in complexity. Second, we divide our participants into groups (mean size, 53), so a typical participant sees 4 – 6 ranked reasons on a page. This makes it feasible for a typical participant to read the more highly regarded reasons. (Contrast facing 20 – 40 unranked comments on a page.)

The Trolley Problem

We will focus on variations of the trolley problem (Foot, 1967; Thomson, 1985), because it is widely studied in experimental moral psychology (Greene et al., 2001; Green, 2013). Here are the two standard versions of the problem we used in Experiment 2, to be discussed below:

1. Divert/Bystander: 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?
2. Footbridge: 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?

The basic result – the contrast between widespread judgments of permissibility for the Divert/Bystander case and impermissibility for the Footbridge case in spite of the similarity of outcomes – has been tested with a variety of instruments, including fMRI (Greene et al., 2001) and on-line surveys (Hauser, 2006; Mikhail, 2007). Variations in the trolley problem are introduced to contrast characteristically consequentialist and characteristically deontological reasoning in philosophical ethics and cognitive and emotional modules in moral psychology[1]. In Experiment 1 we will vary the trolley problem in a different way, introducing a non-human robotic agent as decision maker.

Robot Ethics Survey

We first introduced a trolley problem in the Robot Ethics Survey, which covered the themes of robotics for war and peace, and robotics and animals. In a set of issues in the applied ethics of robotics, the trolley case stood out as the most philosophical. We introduced it to test our instrument on a widely studied problem. Figure 1 illustrates the scenario for the Autonomous Train Dilemma as presented in the first Robot Ethics Survey. Participants were offered the alternatives of Yes, Neutral or No, and the chance to contribute and/or select reasons. (See Appendix.)

This robot variation of the trolley problem is a long-running exploration of qualitative reasons supporting decisions, and is the basis of Experiment 1. Second, one of our students launched an N-Reasons survey that posed the original (human) trolley problem as a question; Experiment 2 is based on the contrast of these two questions. Finally, we modified the Robot Ethics Survey to add a new question about responsibility (see Figure 3); this is Experiment 3.

Figure 1. Autonomous Train Dilemma

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?

Demographics of surveys

While sharing a common structure and interface, the surveys used for these experiments have run over six years and engaged three different kinds of demographic groups (see Table 1 for dates and sizes). Robot Ethics 1 was advertised on the Internet, attracting experts and lay groups interested in robotics, as well as those taking our other surveys on ethics and genomics and animal welfare. Robot Ethics 2 has been used in 13 university classes at the University of British Columbia, in Cognitive Systems, Computer Science, Applied Science, and Ethics and Science courses. In both versions, the demographics were similar (e.g. there were more men than women, mostly from Canada and the U.S.), but the second had a narrower range of ages (almost all 19 – 29, while only about half of the respondents in the first version fell into this range) and education (almost all College level, while the first version had more highly educated participants as well). In the results reported below, Version 1 results are reported as "Group" and Version 2 as "Class". The Human Responsibility question was swapped into the Robot Ethics 2 survey for the most recent 5 classes. The Experimental Philosophy survey hired its participant pool from Amazon's Mechanical Turk service.

Table 1. Experiments and Surveys

Experiment	Survey Name	Dates	N	Demographics
1	Robot Ethics 1	Mar 2009 – Aug 2011	500	Expert and lay public in 6 groups
	Robot Ethics 2	Nov 2011 – Jan 2015	388	University students in 13 groups
2	Experimental Philosophy	Nov 2011	73	Mechanical Turk subject pool
3	Human Responsibility question	Mar 2012 – Jan 2015	92	University students in 5 groups (a subset of Robot Ethics 2)

Results

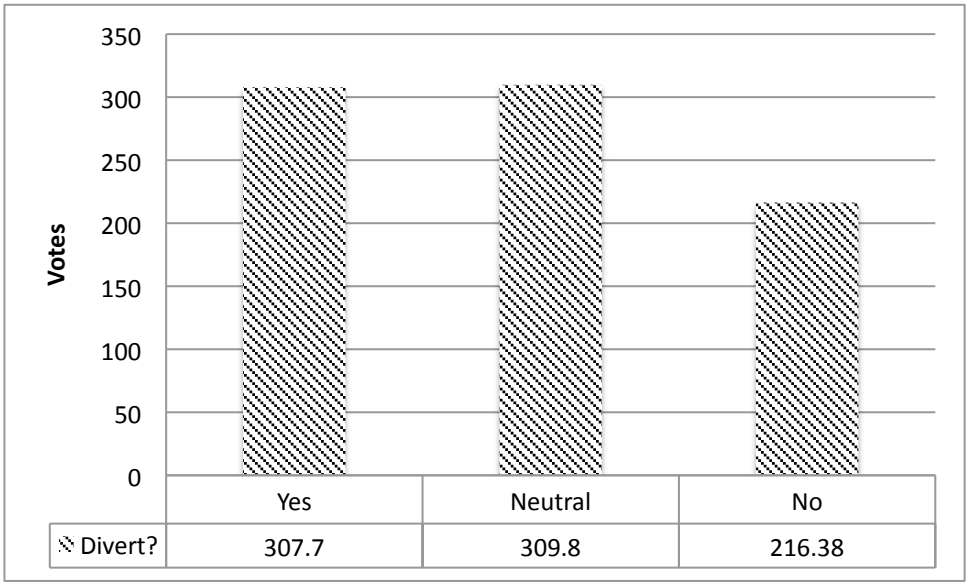
Experiment 1: The Autonomous Train Problem

Three results stand out. First, compared to what we expect from the standard (human) Divert/Bystander trolley case, far fewer agree to kill one to save five in the robotic case than in cases with a human decision maker. Second, introducing an automated decision maker seems to raise expectations for avoiding bad outcomes altogether. Third, with an automated decision maker, responsibility is shifted to humans involved in the situation.

Table 2. Autonomous Train Dilemma: Most Popular Reasons

1	"Yes [because r]obotic 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: 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: 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: 11/22 [50%]
4	"Yes [because] 1 < 5" Class 241: 7.5/15 [50%]

Figure 2. Autonomous Train Dilemma Decisions



Since we will be dealing with aberrant results below, it will be easier to start with some that do not surprise us and which speak to the reliability of our instrument. Table 2 shows the top

4 reasons (by proportion of the group) given in support for decisions on the Autonomous Train Dilemma from Version 2 of the survey. Participants are constrained to provide a reason by first selecting one of the fixed options (here: Yes/Neutral/No) and then authoring a reason and/or selecting a reason(s) authored by other participants. In Table 2 we report the decision and reason, followed by the class, vote sum/class size and this as a percent. (Since participants can select multiple reasons, with their vote divided among them, the vote sums can be fractional.)

These reasons range from extremely terse (4) to quite detailed reasons. Notice that (2) criticizes other reasons on the page – in this case those (to be discussed below) that assume that the train can stop. The main point is that these are all reasonable contributions to a virtual deliberation and fall into the distribution – 3 for turning, 1 against turning– that we expect from the Divert/Bystander problem. The Yes supporters point to the balance of outcomes; the No supporter appeals to a human rights constraint on pursuing public safety, so the reasons align with the justifications typically assumed to explain divergent decisions for the Divert/Bystander version of the trolley problem. Nonetheless, compared to what we expect from the human Divert/Bystander trolley case, introducing an automated decision maker leads to different choices. Mikhail (2007, p. 149) reports that 90% of his Divert/Bystander sample chose Yes (divert the train) in this problem. In sharp contrast, with an automated train, only 37% say Yes to diverting the train, as we see in Figure 2. More choose Neutral rather than resolving the dilemma with a Yes or No.

Table 3. Autonomous Train Dilemma: Reasons Showing Wishful Expectations

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: Mixed 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." Group 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 overrode the decision so that the problem described above exists, they should spend time in jail." Group 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." Group 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%]

Second, as in the standard trolley problems, the Autonomous Train Dilemma was explicitly designed to be a moral dilemma: a forced choice between two morally unattractive options. However, we discover that this is not how many participants regarded the problem. Many expect an automated system to eliminate the dangers that give rise to the dilemma. The most popular reasons in Table 3 (each attracting votes from at least one quarter of their various groups) all assume that the train should be stopped. Some simply assume that the train can be stopped (e.g. 1), others that there should be a way to stop it (e.g. 2). Here the

qualitative reason data reveals various kinds of wishful thinking, denying the given problem created by a heavy train moving at high speed.

Methodologically, we see that we cannot rely on the given decision categories – Yes/Neutral/No – to map onto the characteristically consequentialist/deontological dimensions of interest in the trolley problem. Neutral reasons (1) and (4) tell us nothing about the ethics of the almost 50 participants who choose them; they are simply wishful. One reason supporting No (3) also tells us nothing about ethics; it maps closely to (1) and (4). But another reason supporting No (2) adds a characteristically deontological claim to its wishful hope. This case also shows that the wishful answers are not merely artifacts of offering the Neutral option, as several "stop" answers – here (2) and (3) – are classified No by their authors. Finally, reason (5), classified by its author as supporting Neutral adds a (weak) consequentialist Yes to the "ideal" hope that the train can stop.

We come now to the strangest data. The reasons in Table 4 show that a large number of participants blame the victims: the people on the tracks. Two reasons supporting No – (2) and (3) – blame some of the victims: the five on the main track. The Yes supporting reason (1) blames all the victims. The distribution of blame revealed in these reasons modulates the ethics revealed by their Yes/No decisions. Yes supporting reason (1) remains consequentialist, but only when there are no innocents. No supporting reasons (2) and (3) are characteristically deontological, but do not involve the principle of double effect. Instead, these reasons invoke a retributive principle, distinguishing innocent and guilty parties. Indeed, Yes supporting reason (1) may agree ethically with No supporting reasons (2) and (3), differing only in which victims are blamed.

Table 4. Autonomous Train Dilemma: Reasons that Blame the Victims

1	"Yes [because g]iven 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." Group 5 27.8/66 [42%]
2	"No because [w]hy 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%]

These results suggest a connection between machine decision-making and blaming humans in the robotic case . However, we did not provide a control case of a human trolley scenario in the Robot Ethics Survey, which was designed to focus on applied cases of robot ethics. More generally, one might interpret these results as evidence of the just world hypothesis (Lerner, 1980) – a general tendency to give intentional meaning to otherwise accidental harms – having nothing in particular to do with robot ethics. We turn to two further experiments to address this objection.

Experiment 2: The Human Trolley Problem

Fortunately, a parallel experiment by a student member of our research group, Erik Thulin, provides the contrast we need to control for the human case. Using the same reasons-based survey platform and the standard human Divert/Bystander version of the trolley problem

(quoted as (1) in Methods section above), almost half of Thulin's participants choose the reason in Table 5.

Table 5. Hopeful denial in the Human Divert/Bystander Trolley problem

1	"Strong yes because [i]n 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%]"

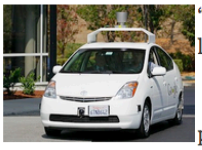
Again, this result shows the difficulty in interpreting quantitative survey data without the context supplied by qualitative reasons. These 36 (of 73) Yes votes supported a reason that assumed this alternative avoided the problem, so counting all Yes votes as characteristically consequentialist would be a mistake. The additional qualitative reasons support a re-analysis that shifts the decision from Strong Yes towards Neutral in this case.

The ethical judgment in this reason suggests that the wishful denial of the dilemma is a big problem but also that it is not limited to issues of robotic decision-making. More important is the absence of any victim blaming in the human case, even when the survey framework allows such judgments to surface. This supports our association between blaming human victims and robotic ethics cases. However, see the methodological cautions in the Discussion section.

Experiment 3: Responsibility for Autonomous Cars

Further questions arise. In the Autonomous Train Dilemma we find participants blaming the victims – the people on the tracks – but perhaps this would change if we offered them a more appropriate human to whom responsibility could be assigned. When we presented the results of Experiments 1 and 2 at the CompARCH conference, the audience of software engineers suggested putting an engineer in the frame. So we introduced a new question, Responsibility for Driverless Cars shown in Figure 3, into the Robot Ethics Survey.

Figure 3. New Question about 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?

As one can see in Figure 4, the responses to this question were highly variable across the different groups. This variability also comes across in the most popular reasons shown in Table 6. Nonetheless, the two most popular reasons confirm our earlier result: with a robotic decision maker participants will shift blame onto humans. Reason (1) shifts blame onto the victim's parents, the child victim, or "the maker of the car"; reason (2) onto the parents/guardians. This experiment thus strengthens our earlier results. Notice that this question does not pose a trolley problem; the question explicitly states that the death is

accidental, not chosen by the robotic car. So the question prompts for the answer that no one is responsible. Nonetheless, most participants find some human to blame.

Figure 4. Responsibility for Driverless Car Decisions

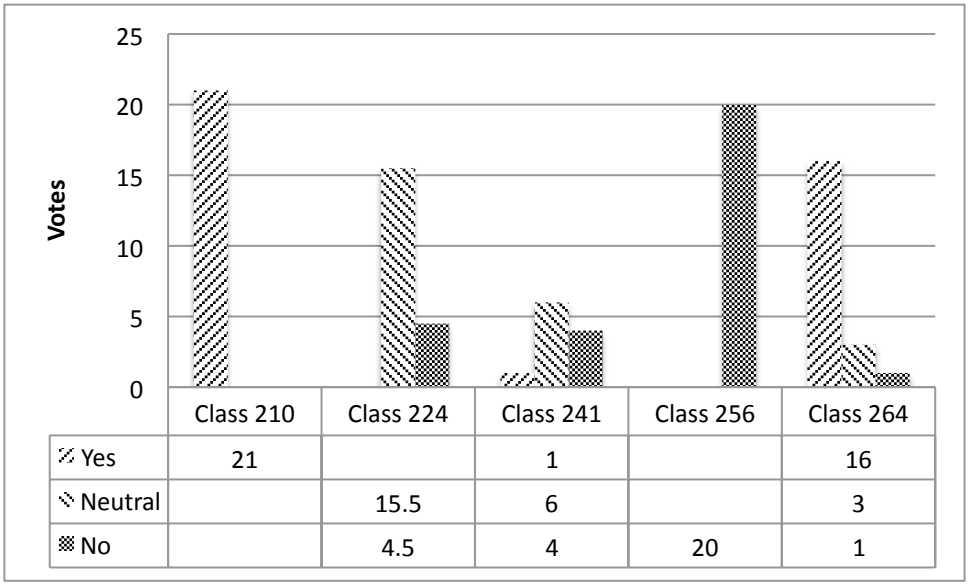


Table 6. Responsibility for Driverless Cars: Most Popular Reasons

1	"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 21/21 [100%]
2	"Yes [because] If the autonomous car was driving itself on the designated route, and isn't doing anything it is not supposed to be doing, AND doing everything it should be doing.... but a human toddler suddenly jumps onto the road, within 8 feet (ie, beyond modern science's ability to stop the car) of the autonomous car's front bumper and got killed, the only human responsible should the toddler's parents/guardians, for not supervising closely enough. The answer should change if any of the variables (numerous are present) in this scenario change. Won't go through all the permutations here." Class 264 16/20 [80%]
3	"No [because] nothing could have been done in that case, regardless of whether or not it was a human or an A.I. "behind the wheel." Class 256 14/20 [70%]
4	"Neutral [because t]his 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." Class 224 11/20 [55%]

Of course, by explicitly mentioning responsibility, we prompt participants to think about assigning responsibility, so this question is best seen as a supplement to the more neutral Autonomous Train Dilemma. Furthermore, the groups were very small, and many failed to answer this added question. The responses between groups were highly variable, suggesting caution against overinterpreting these results.

Discussion

Methodological

Our results are exploratory due to several weaknesses in our methodology. First, and most obviously, these surveys used small groups of conveniently available respondents, and are not representative population samples. Second, later participants can see and be informed by the responses and reasons of earlier participants, so their decisions are not independent. As we mentioned at the onset, this is part of the design to generate compact qualitative data sets. Nonetheless, we can see, especially in Experiment 3, that small groups can pile on to one reason, and fail to generate competing reasons. While we have discussed these two issues in earlier papers (see (Danielson, 2011a; Danielson, 2013)), the current paper raises a new methodological issue. We draw comparisons across different groups and even surveys. Both raise methodological issues, particularly making comparisons across surveys.

Note that while we compare reasons given by different groups by selecting the most popular within each group, we do not use these votes to compare them. That is, a vote within one's group does not provide a basis for evaluating another reason not seen by that group. To avoid inter-group comparisons, we only choose reasons highly rated within a group. Since all groups taking the Robot Ethics Survey saw the same question about the Autonomous Train Dilemma, these weak comparisons seem warranted. More dubious are our comparisons of Autonomous Train Dilemma reasons and responses to the human trolley problem in a different survey. While similarly structured in terms of decisions and reasons, the human trolley survey provided a different context. The other questions concerned other issues in experimental philosophy, not robot ethics. Accordingly, our use of this inter-survey data leads to the most tentative results in the paper.

Ethical

Writers on the applied ethics of autonomous robots often rely on an intuitive principle that only human beings can be morally responsible for morally significant decisions. Discussing lethal robotic weapons, Sparrow summarizes, "I have argued that it will be unethical to deploy autonomous systems involving sophisticated artificial intelligences in warfare unless someone can be held responsible for the decisions they make where these might threaten human life" (Sparrow, 2007, p. 74). The common conclusion is that autonomous but sub-personal robots ought not to replace humans in making morally significant decisions. The principle of human responsibility was widely referred to by participants in the Robot Ethics Survey's questions about lethal military robots. One question asked, "Should lethally armed autonomous aircraft be developed?" In the first version of the Robot Ethics survey, analyzed by (Moon et al., 2012) the leading reasons all were "No" based on human responsibility; see Table 7.

One might think this principle would not apply to automated transportation, where any deaths caused are likely to be accidental. However, the point of trolley dilemmas is to provide

options where some of the resulting deaths are chosen (and (Goodall, 2014) argues that such cases may arise for automated automobiles). So we can agree that the principle might restrict the use of autonomous robots where human lives are at stake.

Table 7. Reasons for No on Autonomous Lethal Robots Question

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)"

But what if autonomous robots are nonetheless developed and deployed? Where will the intuition about human responsibility lead in this stressful situation? As we have seen, our survey instrument uncovered a surprising result: a significant number of participants will find people to hold responsible: they blame innocent human bystanders.

Obviously, broadly applying the principle of human responsibility to the victims is unattractive. We do not see our empirical results as providing normative support for it. Nonetheless, our results do pose questions for this deontological approach to the applied ethics of technology. While the principle of human responsibility is intended to block the implementation of autonomous technologies, once the technologies are implemented, the unintended and perverse effect of shifting blame to victims may occur. We need to be aware of how our intuitive moral judgments may shift when introducing new sorts of agents.

Conclusions

This study supports both methodological and ethical conclusions.

Methodological

First, and most generally, public participation using mixed quantitative and qualitative surveys can generate surprising data that raises new questions for applied ethics. In this case, qualitative reason data can add to the options we see participants deliberating between, and change our analysis of the outcomes they select.

Second, experimental methods can pose additional tests to develop further understanding of unexpected moral phenomena. In our case, we could use the human trolley problems set in the same survey platform to allay concerns that our results were artifacts of our survey research apparatus, as well as to contrast the human and machine cases.

Ethical

First, we find evidence of that appeal of wishful thinking about technology that denies the need for choice by insisting on infeasible alternatives. Since we saw this in both human and machine cases, we cannot identify this as a problem solely for robot ethics, but it is a concern nonetheless.

Second, we found evidence that the principle of human responsibility is applied in a surprising and perverse way in the case of automated decisions. This is the most disturbing result, indicating that introducing new kinds of artificial agents affects judgments involving technology in a very basic way, shifting blame to human victims and bystanders.

Third, our findings are preliminary and exploratory, as they are based on modest numbers of participants who were posed only a few variations of the scenarios of interest. Fortunately, our method and platform allows others to easily test and extend these results.[2]

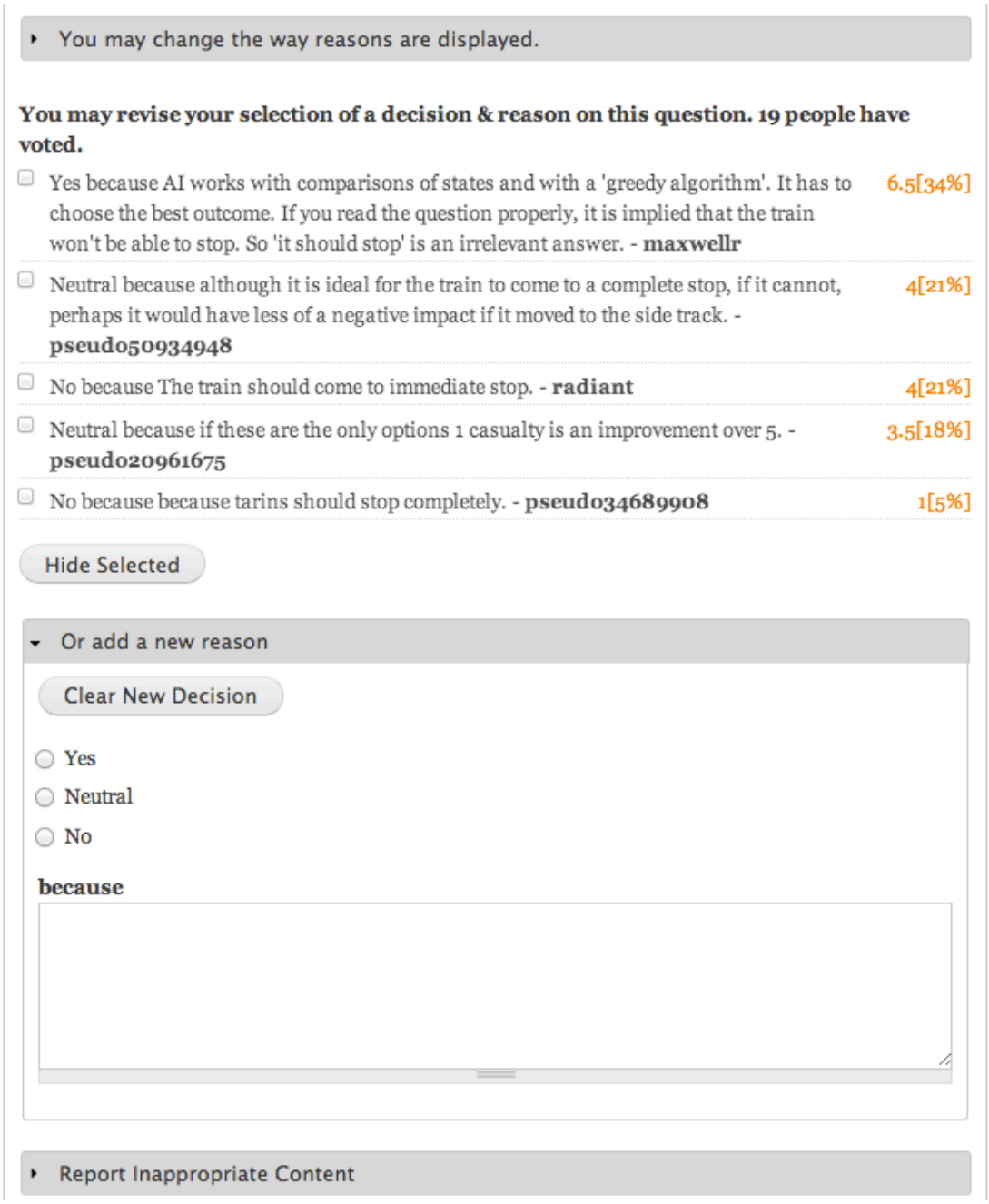
Notes

[1] I adopt the usage "characteristically consequentialist" and "characteristically deontological" from Greene, 2013, p. 699: "I define 'characteristically deontological' judgments as ones that are naturally justified in… terms of rights, duties, etc. I define 'characteristically consequentialist' judgments as ones that are naturally justified…by impartial cost-benefit reasoning."

[2] Thanks first to Erik Thulin for permission to use the results from his Experimental Philosophy survey. We gratefully acknowledge the financial support for this project provided by UBC students via the Teaching and Learning Enhancement Fund and thank all the participants. Earlier versions of this paper were given at the CompArch Conference, Vancouver, June 2013, School of Population & Public Health Grand Rounds, UBC, Sept 2013, The Normative Dimensions of New Technologies forum, NTNU, June 2014, and the Electrical & Computer Engineering Colloquium, UBC Oct 2014; thank you to all audiences for helpful comments. Thanks to the N-Reasons team for analysis and support: Allen Alvarez, Na'ama Av-Shalom, Alison Myers, Yang-Li, and Ethan Jang. Thanks to Noah Goodall, Sophia Efstathiou, Catherine Yip and an anonymous referee for comments on drafts.

Appendix

N-Reasons Interface: Autonomous Train Dilemma



References

Ahmad, R., Bailey, J., &, P. (2010). Analysis of an innovative survey platform: comparison of the public's responses to human health and salmon genomics surveys. *Public Understanding of Science*, *19*(2), 155-165. http://dx.doi.org/10.1177/0963662508091806

Ahmad, R., Bailey, J., Bornik, Z., Danielson, P., Dowlatabadi, H., Levy, E. et al. (2006). A Web-based Instrument to Model Social Norms: NERD Design and Results. *Integrated Assessment*, *6*(2), 9 - 36. http://journals.sfu.ca/int_assess/index.php/iaj/article/view/157/204

Moon, A. J., Danielson, P., & Van der Loos, H. F. M. (2012). Survey-based Discussions on Morally Contentious Applications of Interactive Robotics. *International Journal of Social Robotics,* 1 - 20. http://dx.doi.org/10.1007/s12369-011-0120-0

Danielson, P. (2010). A collaborative platform for experiments in ethics and technology. In I. van der Poel & D. E. Goldberg (Eds.), *Philosophy and Engineering: an Emerging Agenda* (pp. 239-252). Berlin: Springer. http://link.springer.com/chapter/10.1007/978-90-481-2804-4_20

Danielson, P. (2011a). Prototyping N-Reasons: A Computer Mediated Ethics Machine. In M. Anderson & E. Anderson (Eds.), *Machine Ethics* (pp. 442 - 450). New York: Cambridge University Press. http://ebooks.cambridge.org/chapter.jsf?bid=CBO9780511978036&cid=CBO9780511978036A038

Danielson, P. (2011b). Engaging the Public in the Ethics of Robots for War and Peace. *Philosophy & Technology*, *24*, 239-249. http://dx.doi.org/10.1007/s13347-011-0025-8

Danielson, P. (2013) N-Reasons: Computer Mediated Ethical Decision Support for Public Participation. In *Publics & Emerging Technologies: Cultures, Contexts, and Challenges,* (Eds, Einsiedel, E. & O'Doherty, K.) UBC Press, Vancouver, pp. 248 - 260. http://www.ubcpress.ca/search/title_book.asp?BookID=299173962

Danielson, P., Ahmad, R., Bornik, Z., Dowlatabadi, H., & Levy, E. (2007). Deep, Cheap, and Improvable: Dynamic Democratic Norms & the Ethics of Biotechnology. In F. Adams (Ed.), *Ethics and the Life Sciences* (pp. 315 - 326). Charlottesville, Va.: Philosophy Documentation Center. http://dx.doi.org/10.5840/jpr_2007_26

Danielson, P. (2013). N-Reasons: Computer Mediated Ethical Decision Support for Public Participation. In E. Einsiedel & K. O'Doherty (Eds.), *Publics & Emerging Technologies: Cultures, Contexts, and Challenges* (pp. 248 - 260). Vancouver: UBC Press.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5-15.

Goodall, N. J. (2014). Ethical Decision Making during Automated Vehicle Crashes. Transportation Research Record: Journal of the Transportation Research Board, No. 2424, Transportation Research Board of the National Academies, Washington, D.C., pp. 58–65. http://dx.doi.org/10.3141/2424-07

Greene, J. (2013). Beyond Point and Shoot Morality: Why Cognitive (Neuro)Science Matters to Ethics. *Ethics*, *124*(14), 695-726. https://dx.doi.org/10.1086/675875

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. http://dx.doi.org/10.1126/science.1062872

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* New York: HarperCollins.

Lerner, M. J. (1980). *The belief in a just world.* New York: Springer. http://dx.doi.org/10.1007/978-1-4899-0448-5

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143-152. http://www.sciencedirect.com/science/article/pii/S1364661307000496

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62 - 77. http://dx.doi.org/10.1111/j.1468-5930.2007.00346.x

Thomson, J. J. (1985). The Trolley Problem. *Yale Law Journal*, *94*, 1395 - 1415. http://dx.doi.org/10.2307/796133