Recognize Everyone's Interests: An Algorithm for Ethical Decision-Making about Trade-Off Problems

Tobey K. Scharding
Rutgers University

This article addresses a dilemma about autonomous vehicles: how to respond to trade-off scenarios in which all possible responses involve the loss of life but there is a choice about whose life or lives are lost. I consider four options: kill fewer people, protect passengers, equal concern for survival, and recognize everyone's interests. I solve this dilemma via what I call the new trolley problem, which seeks a rationale for the intuition that it is unethical to kill a smaller number of people to avoid killing a greater number of people based on numbers alone. I argue that killing a smaller number of people to avoid killing a greater number of people based on numbers alone is unethical because it disrespects the humanity of the individuals in the smaller-numbered group. I defend the recognize-everyone's-interests algorithm, which will probably kill fewer people but will not do so based on numbers alone.

Key Words: trade-off problems, algorithms, autonomous vehicles, respect for humanity, trolley problem

Advances in technology have intensified the problem of *trade-offs* in business decision-making. In trade-off problems, businesses must decide between two mutually exclusive options, each of which captures a distinctive good that the other does not. Trade-off problems arise in many domains: which organizations should benefit from corporate philanthropy (Muller, Pfarrar, & Little, 2014), what drugs pharmaceutical companies should develop (Lanteri, Chelini, & Rizzello, 2008), when criminal defendants should be detained as they await trial and when they should be released (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017), whether firms should serve applicants' or corporate interests in hiring decisions (Leicht-Deobald et al., 2019), and how technology companies should program their autonomous vehicles (AVs) to respond to unavoidable crashes that pit the interests of one person (or a small number of people) against the interests of a larger number of people (Bhargava & Kim, 2017; Goodall, 2016; Lin, 2016; Millar, 2017). Now that managers are increasingly delegating trade-off problems, along with other business decision-making, to algorithms (Colson, 2019; Martin, 2018; Parmar & Freeman, 2016), the difficulties they present are especially important to resolve. Whereas human decision makers can weigh independently all of the difficulties that arise in particular decision-making contexts, algorithms implement choices that human designers and engineers made previously, without ever having scrutinized the particular difficulties of those decision-making scenarios (Nyholm & Smids, 2016). In this article, I examine

Business Ethics Quarterly 31:3 (July 2021), pp. 450–473. DOI:10.1017/beq.2020.29
Published by Cambridge University Press on behalf of the Society for Business Ethics.

this broad problem by way of the relatively well-developed AV example: addressing how AV-manufacturing businesses should, as a matter of ethics, program AV algorithms to respond to trade-off problems in unavoidable crash scenarios.

I begin by introducing the problem and identifying the difficulties it creates, ethically speaking, in deciding between four candidate algorithms: *kill fewer people*, *protect passengers*, *equal concern for survival*, and *recognize everyone's interests.* These difficulties include the differences between human and algorithmic decision-making and the challenge for human decision-making of resolving which ethical standard should guide algorithm design—best consequences, self-interest, or respect for humanity—when these standards conflict.

Then, I introduce a model for grappling with this problem: what I call the *new trolley problem.* The new trolley problem is based on the classic trolley problem, which addresses clashes in people's intuitions about which ethical standard—for example, best consequences or respect for humanity—applies in three trade-off crash scenarios in which the decision maker can save either few or many pedestrians. The new trolley problem poses a fourth trade-off crash scenario, in which a decision maker can save both a smaller-numbered and a larger-numbered group of pedestrians, but only by self-sacrificing. In this way, the new trolley problem models the kind of decision-making involved in deciding how to program AVs to respond to trade-off crash scenarios in which the AVs can either save many pedestrians (i.e., by sacrificing passengers) or save the AVs' passengers (i.e., at least in some crashes, by killing more people).

Of the four algorithms, I defend recognize everyone's interests as the most suitable algorithm for AVs, ethically speaking, on the ground that it best meets the ethical standards—best consequences, self-interest, and respect for humanity—pertinent to trade-off crash scenarios. In particular, I argue that equal concern for survival and recognize everyone's interests alone show respect for humanity; thus only these algorithms are permitted from an ethical perspective. Of the two algorithms, recognize everyone's interests does (much) more to satisfy the best-consequences standard; thus I argue that it is the most ethical algorithm for AVs. I conclude by discussing other ways in which the decision-making rationale associated with the recognize-everyone's-interests algorithm can be fruitfully applied in a business ethics context.

AUTONOMOUS VEHICLES

In recent years, ethicists have become interested in how AVs should respond to trade-off crash scenarios, in which the loss of life is unavoidable but there is a choice about whose life or lives are lost (Bhargava & Kim, 2017; Johnson, 2015; Lin, 2016; Nyholm & Smids, 2016). In this section, I present an example of a trade-off scenario involving an AV followed by a menu detailing the various algorithms that could direct the AV's response to the scenario. The remainder of the article will be devoted to determining which algorithm should be used, ethically speaking, to respond to such scenarios.

An Ethical Problem for Autonomous Vehicles

As the article's motivating example, consider the following scenario. An AV carrying one passenger suddenly encounters (that is, without time to brake effectively)

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

four pedestrians in the middle of the road. To the left is a single pedestrian and to the right a stalled semitrailer truck. In this trade-off scenario, the AV must either swerve left, killing one pedestrian; swerve right, killing the passenger; or proceed straight, killing four pedestrians.

The AV's algorithm can direct it to do any of these three things. It can also select among the options probabilistically, such that there is, for example, a 1-in-3 chance of doing each of the options or a 1-in-6 chance of doing one of the options and a 5-in-12 chance of doing each of the other two options.

A Menu of Algorithms to Solve the Ethical Problem for Autonomous Vehicles

The possibilities broached in the preceding section suggest various responses, which I use to name the algorithms associated with them: kill fewer people, protect passengers, equal concern for survival, and recognize everyone's interests. Albeit not exhaustive, this list of possible algorithms is thorough to the extent that it includes the most plausible responses to this AV crash scenario.[1] In the following pages, I explain how each of the algorithms would respond to this scenario and which ethical standard the response serves. I briefly defend the standard with the aim of establishing that the algorithm serves an ethical standard. I do not, though, attempt to arbitrate between the standards or determine which is the most important.

Kill Fewer People

Swerving right or left kills fewer people in the sense that only one person dies if the AV swerves right or left but four people die if the AV goes straight. These actions serve the ethical standard of securing the best consequences (Bentham, 1789; Gustafson, 2013). As the possible responses to this AV crash scenario are distinguished by how many lives are lost in each alternative and killing fewer people is the alternative (by definition) in which the fewest number of people die, the kill-fewer-people algorithm appears to secure the best consequences in this scenario.

Protect Passengers

Doing nothing or swerving left protects passengers in the sense that only swerving right kills the passenger. The protect-passengers algorithm serves an ethical standard associated with self-interest (Maitland, 2002; Sidgwick, 1907) to the extent that protecting the passenger serves the passenger's self-interest (though it does not serve the self-interests of the pedestrians). Here I define interests narrowly in terms of avoiding serious harm or death. Giving passengers priority over pedestrians, motorists, and other parties protects passengers' interests in remaining alive and uninjured.

[1] I have eliminated less plausible responses, such as *protect single pedestrians*, so as to conserve space and focus on the most promising possibilities. For a discussion of the reasoning that could support the latter algorithm (though the authors do not themselves support saving the single pedestrian over the other parties in trade-off scenarios featuring a single pedestrian as one of the parties), see Hsieh, Strudler, and Wasserman (2006: 361–62).

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

Equal Concern for Survival

Giving each scenario a 1-in-3 chance of being realized, in turn, shows equal concern for survival in the sense that it gives each person the same chance of surviving the crash. As such, this algorithm serves an ethical standard of respect for the humanity of individuals as having equal moral status (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785). Equal concern for survival serves this standard in the sense that no person receives any priority over any other; all people implicated in the trade-off scenario enjoy the same chances of surviving.

Recognize Everyone's Interests

Finally, giving each of the six people an equal (that is, a 1-in-6) chance of avoiding death while weighing the decision-making alternatives according to how many people would be saved in each alternative creates a 5-in-12 chance that the AV will swerve right, a 5-in-12 chance that the AV will swerve left, and a 1-in-6 chance that the AV will do nothing (that is, proceed ahead, striking and killing the four pedestrians).[2] This probabilistic strategy for selecting among the three alternatives recognizes everyone's interests in the sense that it considers the interests (i.e., in avoiding death) of all people implicated in the scenario. As in the protect-passengers algorithm, I narrowly define people's interests as avoiding serious harm or death. Recognizing everyone's interests serves the ethical standard of respect for the humanity of individuals as having equal moral status (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785) to the extent that each human being weighs equally on the algorithm's selection.

Three Challenges in Selecting an Ethical Algorithm

The human beings who design algorithms must select among these options; it is up to them to ensure that the algorithms direct the AVs ethically. In doing so, the human designers face three challenges, which I discuss in turn.

Human and Algorithmic Decision-Making

First, recent research has noted that AV algorithms operate differently from human ethical decision-making (Nyholm & Smids, 2016). Humans typically make life-and-death decisions (such as whether to swerve left, swerve right, or proceed straight) in the heat of the moment, without consulting with others. Algorithms, by contrast, are produced over time and by many people. Although this difference could be considered problematic for designing algorithms to make ethical decisions (Nyholm & Smids, 2016: 1280–82), I argue that the algorithm design presents an opportunity to improve AVs' decisions vis-à-vis typical human decision-making. Whereas people could regret decisions made in the heat of the moment or regard—after the fact—their decisions as having been unethical (Williams, 1981), AVs have the power to make decisions that human beings can endorse as ethical both before and after they

[2] I provide the calculations that produce these probabilities in the Proportional-Risk Standard section below.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

are carried out. That is to say, people's reflection and debates about what is the ethical thing to do in trade-off scenarios can empower AVs to respond to trade-off scenarios in ways that are better thought out, more stable, and—perhaps—more ethical than those achieved in ordinary human decision-making. In this sense, programming AV algorithms presents challenges for ethics (i.e., ensuring that the algorithms direct the AVs ethically) but also opportunities for ethics (i.e., facilitating more ethical decision-making in trade-off crash scenarios).

Second, whereas human drivers make (very fast) decisions about their own welfare when facing trade-off crash scenarios, AVs do not make decisions when they encounter trade-off crash scenarios. Rather, the AVs implement whatever responses human designers and engineers have programmed them to make when particular sensory inputs are triggered. As such, AVs lack the kind of self-interest that could motivate human drivers to swerve into groups of pedestrians to save themselves. This lack of self-interest is significant in selecting algorithms for AVs to the extent that it shows that AVs need not put passengers' interests first (in the way that human drivers perhaps cannot avoid putting their own interests first). Instead, AVs are free (i.e., their human designers are free) to serve more ethical objectives, all things considered, thus presenting further opportunities for ethical resolutions to trade-off crash scenarios. One of these ethical objectives remains, of course, passengers' self-interest. To the extent that passengers entrust their safety to the AVs in which they travel and perceive these AVs as serving their interests, I assume that the human designers and engineers that program AV algorithms are bound to take passengers' interests into account. To the extent that passengers own the AVs and are legally responsible for them,[3] moreover, it seems reasonable to think of AV programmers as acting on passengers' behalf.

Human Decision-Making and Ethics

Third, AV-manufacturing businesses must resolve what is the ethical action in various trade-off crash scenarios. Because little consensus exists regarding this issue (Bhargava & Kim, 2017; Bonnefon, Shariff, & Rahwan, 2016; Millar, 2017), the challenge such businesses face is daunting. Among other problems, empirical research demonstrates substantial cultural variation in people's responses to trade-off scenarios (Ahlenius & Tännsjö, 2012; Gold, Colman, & Pulford, 2014).

In response to uncertainty about how people should respond, ethically speaking, to trade-off scenarios, several theorists have sought to apply views about how the classic (and very well-developed) trolley problem in ethics (Foot, 1967; Thomson, 1976, 1985) should be resolved directly to the problem of how AVs should respond to trade-off scenarios (Goodall, 2016; Lin, 2016). Other theorists have found problems with using the classic trolley problem in this way (Millar, 2017; Nyholm & Smids, 2016). In the next section, I take a middle ground, arguing that the

[3] How legal liability will be established in AVs is an open question. Most scholars who have addressed this issue favor a program of strict liability (Colonna, 2012; Duffy & Hopkins, 2013; Villasenor, 2014). Some assert that owners are liable (Brodsky, 2016; Duffy & Hopkins, 2013); others hold manufacturers liable (Villasenor, 2014).

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

classic trolley problem is unhelpful but that a modification—the new trolley problem—can and should guide the human designers of AV algorithms.

THE TROLLEY PROBLEM(S)

I begin by explaining why using responses to the classic trolley problem (Foot, 1967; Thomson, 1976, 1985) to address problems in ethics about programming AVs has seemed promising to a number of theorists (Goodall, 2016; Lin, 2016) and identifying a novel shortcoming in using the classic trolley problem in this manner. Then, I explain the seeds of a solution to this shortcoming in a recent revision (Thomson, 2008) to the classic problem. I call this revision the new trolley problem. A successful response to the new trolley problem, according to my analysis, has the resources to guide human designers and engineers in programming AV algorithms to respond ethically to trade-off crash scenarios.

The Classic Trolley Problem

The trolley problem is a well-known philosophical thought experiment that describes three trade-off crash scenarios in which an out-of-control trolley rushes toward five people trapped on a trolley track. In the first trade-off scenario, the trolley's driver must decide whether to steer the trolley onto another track where one person is trapped. In the second scenario, a bystander must decide whether to use a lever at the side of the track to switch the trolley to another track where one person is trapped. In the third scenario, a bystander must decide whether to push a person in front of the trolley to stop it from hitting the five people. The *problem* of the trolley problem is that no single ethical theory matches people's intuitions about what action is ethical in all three scenarios (Thomson, 1985). In particular, many people feel that the decision maker should, ethically speaking, pursue best consequences (i.e., kill fewer people) in the first scenario but not in the third (Thomson, 1985).[4]

Recently, theorists have sought to apply solutions to this thought experiment to how AVs respond to trade-off crash scenarios (Bhargava & Kim, 2017; Goodall, 2016; Lin, 2016; Millar, 2017). This application obviously differs somewhat from the classic trolley problem. Even if they discuss the second and third scenarios, theorists who use trolley-style scenarios to evaluate ethical questions about AVs exclusively focus on what is ethical in the first scenario (Bhargava & Kim, 2017: 6; Goodall, 2016: 57–58; Lin, 2016: 79). In this sense, they neglect the intuition that makes the trolley problem problematic on Thomson's (1985) view, namely, that it does not always seem ethical to act in a way that secures the best consequences (i.e., kills fewer people). Philosophical approaches to the trolley problem are therefore somewhat off point as regards the question that AVs face in trade-off scenarios.

[4] Although I address the trolley problem from a theoretical rather than empirical perspective, experimental philosophers and moral psychologists have collected significant data regarding how people respond to trolley-style trade-off scenarios (Awad et al., 2018; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Lanteri et al., 2008).

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

The New Trolley Problem

In a 2008 article that has been much less noted than Thomson (1985), Thomson reconsiders the classic trolley problem. Most significantly for my purposes, she offers a novel decision-making scenario that resembles, but is more complex than, the first scenario of the classic trolley problem. The novel, more complex scenario compresses the concerns of the classic trolley problem's first and third scenarios, thereby challenging the intuition that it is ethical to kill fewer people in scenarios such as the first scenario of the classic trolley problem. Because of this compression, an account of what it is ethical to do in the novel, more complex scenario seems *on* point as regards AV algorithm design. The scenario focuses on a car driver, Alfred,[5] who faces two turn-offs from a track where five will die: one in which one person will die and one in which the car driver (Alfred) will die. Thomson's novel scenario thus resembles the AV scenario introduced earlier.

In response to the question of what the car driver may do to save the five pedestrians, Thomson (2008) demurs. The moral equivalence of the car driver and the single pedestrian—each is an individual person who has an interest in not being killed—along with the car driver's reasonable unwillingness to self-sacrifice, confounds her (Thomson, 2008). Though she says that she "prefers" the car driver to self-sacrifice, she considers that it might "be fair. . . to flip a coin" to decide between self-sacrificing and killing the single pedestrian (371).

Thomson's (2008) argument suggests, then, that car drivers need not self-sacrifice (i.e., that they need not self-sacrifice on the grounds that self-sacrificing will kill fewer people) when presented with a choice between their lives and other people's lives. Because they need not self-sacrifice, though, Thomson infers that car drivers may not kill other people to achieve ends that they wish to achieve, such as killing fewer people (i.e., avoiding killing five people). Though she appears to conclude that car drivers "cannot decently regard [themselves] as entitled to *not*" kill one pedestrian to avoid killing five pedestrians, if those are the only choices (372), her extended, intricate discussion serves to bring out the ethically ambiguous nature of trade-off crash scenarios. In particular, she appears to move from 1) preferring a kill-fewer-people solution in (scenarios like) the first scenario of the classic trolley problem to 2) indecision for the reason that there is no difference, ethically speaking, between the individuals in the scenarios; all are "without fault" (371).

The extended, intricate discussion, that is to say, suggests that considerations outside of best consequences—including self-interest and respecting the humanity of the individuals involved in trade-off crash scenarios as having equal worth—are relevant to the resolution of this trade-off crash scenario. Thomson (2008) does not provide this solution. Neither killing five people, nor killing one, nor self-sacrificing

[5] Thomson (2008: 370) discusses a car driver instead of a trolley driver to avoid being influenced, argumentatively speaking, by obligations that trolley drivers might bear to protect track workers. Because algorithm-designing and/or AV-manufacturing businesses also wish to avoid that influence, this change further aligns Thomson's concerns with ethical questions about AVs.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

seems like obviously the ethical thing to do on her analysis. Providing a rationale for what is ethical to do in this scenario is what I call the new trolley problem.

SOLVING THE NEW TROLLEY PROBLEM

I begin by explaining why neither the best-consequences nor the self-interest standard is adequate to solve the new trolley problem. Then, I discuss the standard of respect for the humanity of individuals as having equal moral status (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785). I consider two ways in which respect for humanity can play out in the new trolley problem: via an equal-risk standard and a proportional-risk standard. These standards correspond to the equal-concern-for-survival and recognize-everyone's-interests algorithms, respectively; I address the new trolley problem using ethical standards that are separate from the algorithms to allow for the possibility that the question about the car driver (Alfred) differs in key ways from the question about AVs, thus calling for a different solution. Although these standards are not exhaustive, I consider them sufficient as responses to the problem of how to apply respect for humanity in a trade-off decision-making context to the extent that they constitute the most plausible applications.[6] After introducing each ethical standard, I explain how it addresses the new trolley problem and detail possible weaknesses in the standard.

Why Best Consequences and Self-Interest Cannot Solve the New Trolley Problem

Neither the best-consequences nor the self-interest standard convincingly addresses what (unique) action is ethical to take in the new trolley problem. This is because no unique action produces the best consequence overall nor best serves self-interest in this scenario. In the scenario, a car driver chooses between killing four pedestrians, killing one pedestrian, and self-sacrificing (i.e., killing one car driver) (Thomson, 2008). That is to say, the best consequence is to avoid killing four pedestrians, but there are two different ways of doing this (i.e., killing one pedestrian or killing one car driver), and the best-consequences standard does not offer straightforward grounds to decide between them. Similarly, the self-interest-serving solution is to avoid killing the car driver, but there are two different ways of doing this (i.e., killing four pedestrians or killing one pedestrian), and self-interest does not offer straightforward grounds to decide between them.

The Equal-Risk Standard

I call the first interpretation of how respect for humanity can solve the new trolley problem the *equal-risk* standard. This standard is based on—though it differs from what is proposed in—the article from which Thomson (2008) appears to take her recommendation that decision makers flip a coin in such trade-off scenarios: Taurek

[6] Another application of respect for humanity, do not kill, is not helpful in trade-off decision-making circumstances to the extent that these decision-making circumstances are premised on the idea that it is impossible to avoid killing.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

(1977). The equal-risk standard recommends that decision makers flip a coin to decide which party to save in trade-off crash scenarios.

Before setting forth the standard, I note a similarity and difference between Taurek (1977) and the standard I develop to address the new trolley problem. The similarity is that in Taurek's scenario, a decision maker also must decide between saving one and saving five. The difference is that Taurek does not address the permissibility of killing parties in a trade-off crash scenario in which all possible responses involve the loss of life but there is a choice about whose life or lives are lost. Taurek's scenario concerns, rather, six patients who need a life-saving drug. One of the six must have all of the drug to survive. The other five each requires one-fifth of the drug (Taurek, 1977). Taurek himself, then, might forbid equal risk to decide whose life or lives are lost in the new trolley problem even as he permits equal risk to decide which person or people receive a life-saving drug.

This difference is not germane to the present investigation, however, in the sense that I do not apply Taurek's (1977) interpretation of *his* case to the classic trolley problem or the new trolley problem; rather, I seek to extract Taurek's procedure, equal risk as exemplified by the coin flip, and evaluate whether it offers a plausible solution to the new trolley problem. This is the sense in which the equal-risk standard is based in Taurek's procedure but does not adopt it completely.

Taurek (1977) proposes that the decision maker should flip a coin to decide who receives the life-saving treatment: heads, the drug is divided between the five; tails, the one who needs all of the drug receives it. In making this proposal, Taurek does not suggest that decision makers must allow the five people to go without the drug. Rather, he offers reason that 1) decision makers may allow the five to go without the drug instead of allowing one to go without the drug and 2) decision makers may not sacrifice one person's life to save five people's lives on the basis of numbers alone. Taurek argues that the coin-flipping solution, unlike the solution of giving the drug to the five because they are more numerous, shows *equal concern and respect* for the six people (303). By this, I interpret him as meaning that the coin-flipping solution shows respect for the humanity of each of the six people as individuals having equal moral status (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785). Because respect for humanity is a very well-developed ethical standard, I use it to motivate the equal-risk standard rather than the more contentious (Miller, 1998) equal concern and respect.

In this case, the benefit of the drug is not universally available (Taurek, 1977). Accordingly, everyone must bear some risk (1/2) that she will not be treated and will die. From the standpoint of respect for humanity, the number of people who gain access to the drug—one or five—is not crucial. What matters is that the procedure for determining who gains access to the drug shows respect for the humanity of everyone whom the decision affects. Flipping a coin exposes all of the sick people to a 1/2 risk that they will not receive the life-saving drug (Taurek, 1977).

Equal Risk and the New Trolley Problem

In the context of the classic trolley problem, equal risk appears to recommend flipping a coin to decide whether the trolley driver should 1) allow the runaway

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

trolley to carry on the track where five people are trapped or 2) divert the trolley to the turnoff, where one person is trapped. As in the medicine case, flipping the coin exposes all of the trapped people to a 1/2 risk of death. The single person is not sacrificed to save the more numerous group on the basis of numbers alone.

In the new trolley problem, Thomson (2008: 371–72) sometimes advises flipping a coin to decide which party to a trade-off crash scenario will live, albeit only when the choice is between one or another single person. Her coin flip exposes each person to a 1/2 risk of death, in line with the equal-risk decision-making procedure. To show respect for the humanity of people in all three parties to the new trolley problem—the five pedestrians, the single car driver, and the single pedestrian—according to equal risk, decision makers would need to flip a three-sided coin (or the like), thus exposing each person to an equal 1/3 risk of death.

Possible Weakness in the Equal-Risk Standard

So as better to understand the equal-risk decision-making procedure, I next evaluate a key objection to this procedure. As suggested earlier, Taurek holds a different position from virtually everyone who has written on this topic. His decision-making procedure has, accordingly, been extensively criticized: immediately following its publication (Kavka, 1979; Parfit, 1978), in the forty years since then (Barry, 1995; Scanlon, 1998; Unger, 1996), and more recently (Halstead, 2016; Henning, 2015; Segall, 2016). In this section, I focus on equal risk's difficulties in evaluating extreme cases. I then summarize the upshot of this objection for the equal-risk rationale that decision makers may spare one person even when they could spare five instead.

The main problem is as follows. Taurek (1977: 306) says that he would prefer a coin-flipping resolution even in a decision between saving one or saving fifty with the same amount of drug. He does not, though, indicate whether he would retain his position in a 1-to-500 or 1-to-50,000 decision. This is potentially a defect of the view. At some point, it seems absurd not to take the numbers into account (Kamm, 1985; Lubbe, 2008).

Would a defender of the equal-risk standard argue that respect for humanity requires decision makers to flip a coin to decide whether to provide a drug either 1) to one person who needs all of it to survive or 2) to 50,000 people, each of whom needs 1/50,000th of the drug to survive? Taurek could argue that in this case, respect for humanity recommends that the 50,000 people each receive the share of the drug they need to survive—on the grounds that flipping a coin to make the decision gives too little weight to their interests and wrongly overvalues the interest of the one person who needs all of the drug to survive (Hirose, 2004; Kamm, 1985). Then again, he might insist that respect for humanity requires the decision maker to value the interest of the one person alongside those of the 50,000 people. Otsuka (2000), for example, defends Taurek by noting that a coin flip views everyone's claims as equally important. Such controversies suggest the limitations of the equal-risk standard as regards extreme cases.

In sum, the equal-risk approach brings up a compelling point about the ethical importance of considering equally the interests of each person implicated in a trade-

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

off scenario. Such interests are narrowly defined, as previously, in terms of avoiding serious harm or death. Its decision-making procedure, the coin flip, is adept at navigating the question of who to save. Equal risk encounters difficulties, though, in extreme cases. Even if such one-or-a-million scenarios seem unlikely to arise for AVs, they are worth considering in terms of the insight they provide into what principle should govern AVs in trade-off crash scenarios, particularly because some scenarios could involve large pile-ups and significant loss of life. A widely used algorithm's decisions, moreover, would affect millions of lives.

The Proportional-Risk Standard

To address trade-off scenarios in which the numeric consequences appear important (such as those in which a large number of people's lives are pitted against a single life) while meeting the standard of respect for humanity, I turn to the *proportional-risk* standard. Like equal risk, this standard divides the risk of death among the parties who are implicated in the trade-off crash scenario. Unlike equal risk, proportional risk weighs the risk proportionally according to how many people are in each party, such that parties with larger numbers of people stand a lower risk of being struck and killed.

In the first scenario of the classic trolley problem, for example, proportional risk recommends that, instead of 1) diverting, or not diverting, a runaway trolley set to kill five to a turnoff where it will kill one or 2) flipping a coin to decide what to do, the trolley driver should 3) use a probabilistic mechanism to make the decision. The probabilistic mechanism could involve rolling a six-sided die, spinning a wheel of fortune partitioned into six sections, drawing lots from a choice of six, or the like. The chances of survival are divided among the people implicated in the trade-off crash scenario such that, when *N* number of people are implicated in the trade-off scenario, each individual receives a 1/*N* chance of surviving the scenario. When a risk of death affects a group of people in the same way, such as the five people toward whom an out-of-control trolley hurtles, proportional risk sums their chances of survival. In the trolley-driver scenario of the classic trolley problem, then, the probabilistic mechanism creates a 1/6 probability of letting the trolley continue on its present track, thus sparing the single person on the turnoff, and a 5/6 probability of switching the trolley to the turnoff, sparing the five people on the present track. According to this solution, each person has the same probability of being spared: 1/6. Because each of the five people on the trolley track is saved whenever any one of them is, however, their effective probability of being spared is 5/6.

Several theorists have elaborated closely related views (Kamm, 1985; Timmerman, 2004), though none have applied them to the new trolley problem or ethical issues concerning AVs. Like Taurek (1977) with respect to the equal-risk standard, these theorists are important inspirations and sources for the proportional-risk standard, but I do not claim that they would endorse the uses to which I put proportional risk, the evaluations I draw from it, or even my conception of the standard itself.

One of the chief appeals of proportional risk is its ability to mediate between quantitative and nonquantitative standards (Cureton, 2009; Lazenby, 2014), such as

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

Table 1: Best Consequences, Proportional Risk, and Equal Risk

Trade-off scenario	Best consequences expected value	Proportional risk expected value	Equal risk expected value
Kill 1 or 5	5	4.33	3
Kill 1 or 10	10	9.18	5.50
Kill 1 or 100	100	99.01	50.50
Kill 1 or 50,000	50,000	49,999	25,000.5

best consequences and respect for humanity. Proportional risk captures values associated with best consequences to the extent that it takes the number of people whose lives are threatened into account. Taking the number of people whose lives are threatened into account allows proportional risk to produce consequences whose expected values resemble those associated with the best-consequences standard. For the expected values associated with the best-consequences, proportional-risk, and equal-risk approaches to the first scenario of the classic trolley problem, for example, see Table 1.

In the classic trolley problem, best consequences saves five with a probability of 1 (thus offering an expected value of 5). Proportional risk, by contrast, has an expected value of 4.33 (saving five with a probability of 5/6 and saving one with a probability of 1/6). Equal risk, in turn, has an expected value of 3 [(5 × 1/2) + (1 × 1/2)]. Proportional risk is thus significantly closer to best consequences than equal risk, particularly as the number of people implicated in the trade-off scenario increases.

Proportional risk thus seems able to avoid the worry about extreme cases that stymied the equal-risk approach (Gertken, 2016; Lawlor, 2006; Lazenby, 2014). In the 50,000-to-1 decision considered in the foregoing section, for example, best consequences clearly requires decision makers to choose save 50,000: doing so has 50,000 times better consequences than saving one. Under equal risk, there is an equal chance that each track will be chosen. Under proportional risk, though, there is a 99.8 percent chance that saving 50,000 will be chosen. Proportional risk is thus vanishingly close to best consequences in extreme cases.

Proportional risk also goes some distance to satisfy the standard of respect for humanity. Like equal risk, proportional risk randomizes the risk of who lives and who dies, such that neither of an undesirable set of alternatives (killing many people, killing one person, self-sacrificing) is chosen on the basis of numbers alone. Proportional risk demonstrates respect for the humanity of each of the people involved in a trade-off crash scenario to the extent that each person is considered (and has the same numeric importance) in its quantitative decision-making process. So, when six people are affected by a decision, each counts for 1/6 under proportional risk.

Proportional Risk and the New Trolley Problem

In the new trolley problem, a car driver faces a decision between allowing a car whose brakes have failed to continue traveling on its present course, in which it is set

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

to strike and kill four[7] people, to steer the car onto one turnoff, where it will kill one person, or to steer the car onto another turnoff, where it will kill the car driver.

Proportional risk takes the interests of all those affected—the car driver and the five pedestrians—into account. This approach allocates to the car driver the same chance of survival (1-in-6) as all of the other people affected in the trade-off scenario. In the new trolley problem, this means that the single pedestrian and the car driver each enjoys a 1-in-6 chance of survival, while the four pedestrians, whose chances of survival are summed, enjoy a 4-in-6 (that is, a 2-in-3) chance of survival.

Each of these chances of survival can be enjoyed in two ways: the single pedestrian survives when the car hits the four pedestrians *or* the blunt object (thus proportional risk mandates a 1/12 probability that the car will hit the four pedestrians and a 1/12 probability that the car will hit the blunt object). The four pedestrians survive when car hits the single pedestrian *or* the blunt object (so there is a 1/3 probability that the car will hit the single pedestrian and a 1/3 probability that the car will hit the blunt object, under proportional risk). The car driver survives when the car hits the single pedestrian *or* the four pedestrians (thus entailing a 1/12 probability that the car will hit the single pedestrian and a 1/12 probability that the car will hit the four pedestrians for proportional risk). Aggregating those probabilities for the three possible responses to this trade-off crash scenario, proportional risk recommends that the car hit the single pedestrian with a 5/12 probability, that the car hit the four pedestrians with a 1/6 probability, and that the car hit the blunt object, killing the car driver, with a 5/12 probability.

Possible Weakness in the Proportional-Risk Standard

This approach to resolving the new trolley problem thus imposes significant risk of death on all parties. Unlike the best-consequences standard, it does not spare the group of four. Unlike the self-interest standard, it does not spare the car driver. Unlike equal risk, some parties (i.e., the car driver and the single pedestrian) face a risk of being hit that exceeds both that of one of the other parties (i.e., the group of four) and the risk of death evenly divided among the parties. In this section, I consider shortcomings in this approach from the perspectives of respecting the humanity of individuals as having equal moral status (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785), of best consequences (Bentham, 1789; Gustafson, 2013), and of self-interest (Maitland, 2002; Sidgwick, 1907).

Regarding the objection from respect for humanity, proportional risk counts each person at the same weight but does not give everyone involved in a trade-off scenario an equal chance of surviving that scenario, as equal risk does. The disparate chances of survival suggest that proportional risk does not value all lives

[7] In my application of the proportional-risk decision-making procedure to the new trolley problem, I use the number of pedestrians, four, associated with my four-pedestrians case rather than the number, five, in Thomson's (2008) discussion because the calculations are easier to relate to the classic trolley problem if the number of people affected by the decision remains six. I do not believe that this minor change adversely affects my argument in any way.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

equally, which respect for humanity (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785) appears to require on the ground that all human lives have equal status, ethically speaking.

To set the stage for considering this objection, recall the first scenario of the classic trolley problem. That scenario involves six people; proportional risk gives each person, accordingly, a weight of 1/6. Five of the people are affected in the same way —all five will die or all five will be spared—thus proportional risk sums their probabilities of being spared such that the group (and, effectively, each of them) enjoys a 5/6 probability (1/6 × 5) of being spared. In the classic trolley problem, then, proportional risk allows the one person on the turnoff only a 1/6 chance of survival while giving each of the five people on the trolley track (for practical purposes) a 5/6 chance of survival. Under equal risk, by contrast, all six people experience an equal risk of death: 1/2. Thus, evidently, proportional risk fails to demonstrate respect for humanity in the sense of valuing all lives equally.

To rebut this objection, I consider two arguments for the claim that summing similarly situated parties' chances of survival meets the standard of respect for humanity: Hsieh et al. (2006) and Scanlon (1998). First, Hsieh et al. (2006) argue that decision makers should sum similarly situated parties' chances of survival because doing so allows them to treat parties more equally than the alternative (i.e., giving each party an equal chance of survival in the manner of the equal-risk approach discussed herein). Decision makers have a scarce good—survival—to allocate, and they should, on the basis of the "right to an equal share," allocate it as equally as possible (Hsieh et al., 2006: 353). When six people are implicated in a trade-off scenario, as in the trolley-driver scenario of the classic trolley problem, allocating the good of survival to the five people rather than to the one person does the most to share that good equally among the affected people.

Scanlon (1998), in turn, argues that summing similarly situated people's chances of survival allows each person's life to weigh on the decision. When fewer lives will be lost in one of the alternatives, decision makers should, ethically speaking, select that alternative (Scanlon, 1998). Decision makers sum people's chances of survival with similarly affected others, that is, *in order* to treat each person as having "the same moral force" (Scanlon, 1998: 232).

In light of this discussion, it seems to me that respect for humanity is ambivalent with respect to equal risk and proportional risk. Respect for humanity, that is to say, offers grounds both to support and to criticize these standards. This ambivalence need not discredit respect for humanity as an ethical standard in this regard. It shows, instead, that respect for humanity is not sufficient to resolve what is the ethical thing to do in trade-off scenarios.

Next, I consider two objections from the standpoint of best consequences. The first concerns 1-or-50,000-style trade-off scenarios in which proportional risk (randomly) selects the single person to be saved. The second worry addresses trade-off crash scenarios in which a car driver, such as Alfred, faces a trade-off crash scenario in which a single pedestrian appears suddenly in front of the car and proportional risk recommends (i.e., as a result of its random, probabilistic process)

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

that he swerve from that path to strike and kill a group of four pedestrians at the side of the road.

Regarding the trade-off crash scenario in which proportional risk (randomly) selects 50,000 people to die over a single person, I respond that such actions are essential to the proportional-risk approach, not an objection to the approach. Although all parties affected by this action, including the single person who was saved, would presumably experience great regret at the loss of so many lives (Williams, 1981), the tragic nature of the decision does not make it wrong. To evaluate the action, one must correctly construe the alternatives that decision makers face. On one hand, decision makers can demonstrate respect for the humanity of each person involved in a trade-off scenario by including each person in probabilistic calculations about which party to save, which sometimes results in larger-numbered parties being killed. On the other, decision makers can always save the larger-numbered party, which means that individuals in smaller-numbered parties are discounted from the decision-making process. Given these decision-making circumstances, along with the ethical importance of respect for humanity and the fact that even consequentialists do not always require decision makers to select the optimal outcome (i.e., so long as the outcome selected meets other standards) (Sinnott-Amstrong, 2019), I hold that decision makers should, ethically speaking, always respect everyone's humanity and occasionally permit a suboptimal outcome.

Next, I consider the objection that in a trade-off crash scenario in which Alfred's car is headed toward a single pedestrian and can only avoid hitting the singleton by swerving in such a way that it will hit and kill five pedestrians, proportional risk gives Alfred a 1/6 chance of swerving toward the five. This assignment of probabilities could seem counterintuitive to the extent that the car is (unintentionally) directed toward the single pedestrian and requires an intentional action (the driver's swerving) to redirect it toward the four pedestrians. The redirection thus violates the best-consequences standard to the extent that it makes it less likely that the best-consequences standard in these situations (kill fewer people) will be met.

This objection is worrying to the extent that it directly pits the best-consequences standard against respect for humanity, thus making it impossible to meet both of these important ethical standards. In such situations, it seems to me that respect for humanity plausibly acts as a constraint on best consequences —on the same rationale as the 1-or-50,000 example discussed earlier. Decision makers can either respect everyone's humanity by including every person in probabilistic calculations about which party to save, which sometimes results in larger-numbered parties being killed, or they can always save the larger-numbered party on the basis of the numbers, which means that individuals in smaller-numbered parties are excluded from the decision-making process on the basis of the size of their group alone, thus discounting their humanity. According to proportional risk, each person implicated in the trade-off crash scenario receives an equal chance of surviving the crash even as proportional risk makes it much more likely (a probability of 5/6) that the car will *not* swerve to kill the

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

larger-numbered group. In this sense, I argue that proportional risk does more to diffuse this objection (by doing more to satisfy both of these important ethical standards) than the alternatives I have considered.[8]

Relatedly, this scenario presents an opportunity to consider if one of these standards has priority over the other or if the scenarios enjoy equal importance, ethically speaking. This 1-or-5 scenario, along with proportional risk's response to it, suggests that respect for humanity and best consequences are at a parity (or near parity) in terms of ethical importance in trade-off scenarios like these and that infrequently permitting a larger-numbered group to be killed to save a smaller-numbered group is ethical as a result of this near parity.

Finally, I consider an objection from the standpoint of the self-interest standard. When one (or more) of the parties involved in a trade-off crash scenario is large numbered and the car driver is a single person, proportional risk subjects the car driver to a greater risk of death than does equal risk. In the new trolley problem, for example, equal risk gives the car driver a 1/3 risk of death, but proportional risk gives the car driver a 5/12 risk of death. This objection offers the opportunity to bring self-interest into the hierarchy of ethical standards begun in response to the previous objection. If self-interest should, ethically speaking, weigh equally or more heavily on decision-making than respect for humanity and best consequences, then equal risk's greater allowance for self-interest (i.e., in many cases) gives it an advantage over proportional risk.

In response, I argue that the importance of self-interest, ethically speaking, is well captured by respect for humanity. One of the reasons that it has been worthwhile to address my AV scenario via an extended excursion into the new trolley problem is that Thomson's (2008) discussion shows that adding self-interest to a trade-off crash scenario similar to the first scenario of the classic trolley problem serves to bring out the importance of respect for humanity in these scenarios; that is, the importance of self-interest, ethically speaking, is that the car driver (Alfred) is permitted to take his own interests into account in trade-off crash scenarios on the ground that he enjoys equal moral status as the five pedestrians whom he wishes to avoid killing. The car driver's equal humanity does not entitle him to a greater probability of being spared than any other party, however; thus I conclude that equal risk's greater allocation for self-interest does not give it an edge over proportional risk, ethically speaking.

[8] As a side note, I do not think that this argument commits me to endorse proportional risk as the right decision-making procedure in an inverted version of Thomson's (1985) second scenario in the classic trolley problem (such that a trolley barrels toward one person and a bystander must decide whether to pull a lever turning it to kill five people). In such a scenario, the bystander faces different decision-making circumstances, ethically speaking, from the car driver. In particular, the bystander must choose between letting one die and directly killing five, whereas (following Foot's [1967] argument that the trolley driver in what I have been calling the first scenario of the classic trolley problem kills whether or not she turns the trolley) the car driver directly kills in either case. It is not clear from the analysis provided herein that proportional risk is the right decision-making procedure in decisions between killing the members of a larger-numbered group and letting the members of a smaller-numbered group die.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

Selecting a Standard for the New Trolley Problem

There are no perfect solutions to trade-off scenarios by their very nature. The task of business ethicists as regards these scenarios, then, is to seek well-thought-out, stable solutions that meet relevant ethical standards. Thomson's (2008) discussion of the new trolley problem shows that best consequences (Bentham, 1789; Gustafson, 2013), self-interest (Maitland, 2002; Sidgwick, 1907), and respect for humanity (Arnold & Bowie, 2003; Hill, 1992; Kant, 1785) are the relevant standards for trade-off crash scenarios. In the three-party decision associated with the new trolley problem, neither a best-consequences-based approach, such as kill fewer people, nor a self-interest-based approach, such as protect passengers, selects a unique action. In a two-party decision, such as the first scenario of the classic trolley problem altered to pit Alfred directly against five pedestrians, I infer that best consequences and self-interest are also inadequate. Neither self-sacrificing to save the five nor killing five to save oneself appears ethical in light of Thomson's (2008) discussion of the new trolley problem.

That leaves us with the two respect-for-humanity-based approaches, equal risk and proportional risk. Of these, the foregoing discussion demonstrates that proportional risk is a more ethical way of addressing trade-off crash scenarios than equal risk for the reason that proportional risk does (so much) more than equal risk to meet the standard of best consequences. (As is obvious from their basis, both approaches demonstrate respect for humanity, thus also meeting the standard of self-interest.) Despite the significant risk of death to all parties under the proportional-risk approach, then, I argue that it is the best of the approaches examined herein from an ethical perspective. Even as it imposes risks of harm on all parties, proportional risk is one of only two approaches that satisfies respect for humanity, and it does the second best job of meeting the ethical standards of best consequences, which, as illustrated in Table 1, is quite a good job. The fact, then, that this standard imposes risks of harm on all parties is undesirable but appears unavoidable given the highly unfortunate—even "tragic" (Williams, 1972)—nature of this trade-off scenario.

A PROPORTIONAL-RISK ALGORITHM FOR AUTONOMOUS VEHICLES

Four alternatives for an ethical, practical algorithm for AVs have been introduced: kill fewer people, protect passengers, equal concern for survival, and recognize everyone's interests. Each algorithm serves an ethical standard: best consequences, self-interest, and respect for humanity. In the previous section, I showed that a respect-for-humanity-based solution, proportional risk, is the most ethical way of addressing trade-off crash scenarios involving cars and human drivers. Proportional risk uses the same probabilistic decision-making procedure as the recognize-everyone's-interests algorithm: both assign each of the N people implicated in a trade-off crash scenario an equal, $1/N$ chance of avoiding death and sum the chances of similarly situated people (i.e., people who survive the trade-off scenario as a result of the same action). Before inferring that the respect-everyone's-interests algorithm is the most ethical for AVs, I consider differences between human decision-making and the decisions associated with AVs. In particular, I question whether AVs'

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

preprogramming by human designers and engineers means that AVs should, ethically speaking, respond differently to trade-off crash scenarios as compared with human drivers. I argue that the differences between human decision-making and the decisions associated with AVs are not relevant to this question and select the recognize-everyone's-interests algorithm as the most ethical for AVs. Then, I consider whether this selection is an "objective" recommendation (Bhargava & Kim, 2017: 7-8) or whether it is only part of an overall decision-making process for AV-manufacturing firms. I argue that recognize everyone's interests is (i.e., strives to be considered as) an objective recommendation for how AVs actually should respond to trade-off scenarios, ethically speaking.

In the new trolley problem discussed earlier, the driver herself decides how to respond to the trade-off crash scenario; proportional risk is a strategy for guiding how she should, ethically speaking, respond. In the AV scenario, by contrast, the AV responds as a result of preprogramming by human designers and engineers who do not confront the particular trade-off crash scenario and are not present in the AV. Whereas human drivers seem permitted, ethically speaking, to consider their own interests in such cases, algorithm designers' self-interests are not implicated in the same way. This could mean that algorithm designers should not, ethically speaking, consider self-interest (i.e., either their own interests or passengers' interests). Not being bound to consider self-interest, though, could entail that algorithm designers should, ethically speaking, select the algorithm that produces the best consequences.

This problem is quite interesting, as algorithm designers do appear to be more able than human drivers to select ethical ways of responding to trade-off crash scenarios, instead of being unavoidably motivated by self-interest, as broached previously.[9] Algorithm designers' greater ability to choose ethical responses does not mean, though, that they are free to discount respect for humanity and favor best consequences. If my argument herein has been successful, respect for humanity and best consequences are at a parity in trade-off scenarios; thus algorithm designers and engineers are under the same obligation to include respect for humanity in their decision-making (e.g., by selecting equal concern for survival or recognize everyone's interests[10]) as human drivers like Alfred.

Human algorithm designers' and engineers' evaluations of proportional risk and equal risk do seem different from the evaluations of human car drivers, discussed earlier, however. In particular, the recognize-everyone's-interests algorithm appears

[9] This issue also implicates the issue of whether algorithmic responses to trade-off crash scenarios directly kill the people involved in those scenarios or if the algorithms' responses amount to letting those people die, a question central to Thomson's (1976) examination of the classic trolley problem. Again (as in note 8), following Foot's (1967) argument that the trolley driver in what I have been calling the first scenario of the classic trolley problem kills whether or not she turns the trolley, I surmise that the algorithmic response directly kills the people implicated in trade-off crash scenarios.

[10] Given that self-interest is not a concern in algorithm design, I note that proportional risk's and recognize everyone's interests' assignments of lower probabilities to self-interest (i.e., in most cases) is not a reason to prefer equal risk (equal concern for survival) to proportional risk (recognize everyone's interests) in programming AVs.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

to require significantly greater computing capability than the equal-concern-for-survival algorithm. Whereas recognize everyone's interests requires a means of comprehending the extensiveness of risk (e.g., the number of pedestrians involved in trade-off crash scenarios, whether the objects involved in trade-off scenarios are impenetrable), equal risk only needs to calculate how many different responses the AV could possibly take—a much simpler task.

In response, I assume that the state and trajectory of technology involving AVs renders plausible the computational requirements associated with recognize everyone's interests. Some AV-manufacturing companies use logic-based algorithms, for example, which are sets of rules that are explainable and transparent (Arrigoni et al., 2016). The recognize-everyone's-interests algorithm seems technically feasible under a logic-based system. Even granting this technical feasibility, however, I acknowledge that the recognize-everyone's-interests algorithm introduces greater complexity and uncertainty over the equal-concern-for-survival algorithm. According to my analysis, this greater complexity and uncertainty does not challenge the ethical argument offered herein—unless it were the case that the greater complexity and uncertainty associated with recognize everyone's interests means that it produces a lower expected value for saving lives than equal concern for survival, which I assume not to be the case. Thus I conclude that the new trolley problem is sufficiently similar to AV trade-off crash scenarios that a solution to the former (viz., proportional risk) can be used to select the most ethical algorithm for AVs, namely, recognize everyone's interests.

A more challenging question is what business ethicists are to make of this selection. In particular, my defense of the proportional-risk standard is intricate, touching on many different ethical standards; its selection as the solution to the new trolley problem rests on my argument that it meets those standards better than any alternative. In this sense, it could be unclear whether the algorithm that proportional risk selects, recognize everyone's interests, is a candidate for the final word on this matter or is simply another alternative adding to a general conversation.

In their recent research, Bhargava and Kim (2017) offer a way of clarifying this issue. They distinguish between *subjective* and *objective* oughts with respect to trade-off crash scenarios in AVs—between, that is, how individual people believe that AVs should respond to trade-off crash scenarios, ethically speaking, and how AVs actually should respond (assuming that there is a correct answer to that question) (Bhargava & Kim, 2017).

Whereas business ethicists could be persuaded by the argument offered herein and develop a subjective belief that the recognize-everyone's-interests algorithm should guide AVs' responses to trade-off crash scenarios (or could fail to be persuaded and develop a subjective belief that a different algorithm should be used), the recognize-everyone's-interests algorithm, and the proportional-risk standard upon which it depends, represents an effort to develop an objectively correct account of how AV-manufacturing businesses should respond to trade-off crash scenarios in Bhargava and Kim's (2017) sense. Proportional risk is more complicated than ordinary human decision-making, allowing for more precise considerations of the many factors at play in trade-off scenarios; it decides what to do in a manner that is more

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

balanced and impartial than ordinary decision-making. I thus submit recognize everyone's interests as a candidate for how AVs actually should respond to trade-off scenarios, ethically speaking, to the extent that this algorithm represents the best-thought-out, most stable, and—in the end—most ethical alternative. Whether I have succeeded in this argument I leave for my fellow business ethicists to decide.

CONCLUSION

In this article, I posed a question for business ethicists about how AVs should respond, ethically speaking, to trade-off crash scenarios. I considered four candidate algorithms: kill fewer people, protect passengers, equal concern for survival, and recognize everyone's interests. To evaluate these algorithms, I considered following the recent trend in business ethics of examining algorithmic responses to trade-off crash scenarios in light of the classic trolley problem. Owing to a previously unanalyzed shortcoming in these uses of the classic trolley problem, I argued that it is not helpful in evaluating which algorithms (if any) are ethical. I then introduced a novel form of the problem, the new trolley problem, that corrects this shortcoming. The new trolley problem seeks a rationale for why AVs may not kill a smaller number of people to save a larger number of people on the basis of numbers alone (i.e., without showing due respect for the humanity of the individuals implicated in trade-off crash scenarios). After considering two strategies for addressing the new trolley problem, equal risk and proportional risk, I argued that proportional risk does the most to meet the various standards involved in trade-off crash scenarios. Proportional risk, in turn, selects the recognize-everyone's-interests algorithm.

The success of my proposal depends both on issues specific to AVs, such as their technical abilities to perceive the numbers and locations of pedestrians and devise probabilistic responses to relevant hazards, and on its usefulness in addressing other kinds of trade-off problems in business. In the remainder of the conclusion, I discuss several ways in which the recognize-everyone's-interests algorithm can be applied.

First, recognize everyone's interests would be particularly useful in helping businesses to address aspects of their decision-making related to the numbers of people affected by those decisions. As regards questions that are structurally similar to the AV question, such as whether pharmaceutical companies should produce drugs treating illnesses that affect a large number people or drugs treating illnesses that affect a smaller number of people (Lanteri et al., 2008), recognize everyone's interests would make its recommendation probabilistically: weighted (other considerations being equal) according to the number of people affected.

Recognize everyone's interests could also play a role in trade-off decisions that are structurally different from the AV case. Take as an example the algorithms that businesses are increasingly using to make hiring decisions, as discussed by Leicht-Deobald et al. (2019). Many businesses prefer algorithms that produce false negatives (occasionally failing to recommend for hire people who are well qualified) over false positives (occasionally recommending for hire people who are not well qualified). Recognize everyone's interests can address the problem that a false negative-producing algorithm discounts applicants' interests (i.e., the well-qualified

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

applicants who are denied employment opportunities) in being hired for positions for which they are well qualified. A recognize-everyone's-interests tweak to a hiring algorithm could give rejected applicants a chance of being hired at a probability determined by the number of qualified applicants who are rejected divided by the total number of applicants. A similar strategy could be applied to the algorithms used to decide whether criminal defendants should be detained or released into the community as they await trial, as set forth by Corbett-Davies et al. (2017). In the false negative-producing algorithms that public safety-loving communities prefer, a recognize-everyone's-interests tweak could give detained defendants a chance of being released according to the number of nonviolent offense-committing defendants who are detained divided by the total number of defendants.

Recognize everyone's interests can even play a role in routine business trade-offs, such as deciding which persons and/or organizations should benefit from corporate philanthropy, as discussed by Muller et al. (2014). In cases in which businesses must choose between two worthy causes, one of which supports a large number of people and the other of which supports a smaller number of people, recognize everyone's interests can help the business to decide which group to benefit. In each of these possible applications, the recognize-everyone's-interests strategy acknowledges that the numbers do matter but are not the first concern—humanity is the first concern—and puts first things first.

Acknowledgements

I am grateful to many persons and institutions for comments, guidance, and support as I worked on this article. I had the privilege of presenting earlier versions of these ideas at the 2018 Society for Business Ethics annual meeting in Chicago, Illinois; the 2018–19 Program on Topics in the Philosophy of Law, Politics, and Economic Justice at the University of North Carolina–Greensboro; the 2019 Big Ten and Friends Paper Workshop held at Rutgers Business School–Newark and New Brunswick; the 2019 Academy of Management annual meeting in Boston, Massachusetts; and the 2019–20 Management and Global Business Research Seminar at Rutgers Business School–Newark and New Brunswick. I would like to thank Wayne Eastman, Robert Hughes, and Danielle Warren for valuable written comments. Grateful acknowledgement is also given to Ted Baker, Michael Barnett, John Cantwell, Matthew Caulfield, Petra Christmann, Joanne Ciulla, David Dick, Heather Gert, Julian Jonker, Kevin Kolben, Terri Kurtzburg, Daniel Levin, Kristen Martin, Michael Matteson, Sara Parker Lue, Gary Rosenkrantz, Michael Simonetti, and many others for constructive critical comments in discussion. Finally, I wish to thank Bruce Barry and three anonymous reviewers for extraordinarily helpful, thoughtful, and incisive feedback on earlier versions of this article.

REFERENCES

Ahlenius, H., & Tännsjö, T. 2012. Chinese and Westerns respond differently to the trolley dilemmas. *Journal of Cognition and Culture*, 12: 195–201.

Arnold, D. G., & Bowie, N. E. 2003. Sweatshops and respect for persons. *Business Ethics Quarterly*, 1(2): 221–42.

Arrigoni, S., Cheli, F., Manazza, S., Gottardis, P., Happee, R., Arat, M., & Kotiadis, D. 2016. *Autonomous vehicle controlled by safety path planner with collision risk estimation*

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

coupled with a non-linear MPC. Proceedings of the 24th symposium of the International Association for Vehicle System Dynamics, Graz, Austria: 199–208. Boca Raton, FL: CRC Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.

Barry, B. 1995. *Justice as impartiality*. New York: Oxford University Press.

Bentham, J. 1789/1961. *An introduction to the principles of morals and legislation*. Garden City, NY: Doubleday.

Bhargava, V., & Kim, T. W. 2017. Autonomous vehicles and moral uncertainty. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence*: 5–19. New York: Oxford University Press.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science*, 352(6293): 1573–76.

Brodsky, J. S. 2016. Autonomous vehicle regulation: How an uncertain legal landscape may hit the brakes on self-driving cars. *Berkeley Technology Law Journal*, 31(2): 851–878.

Colonna, K. 2012. Autonomous cars and tort liability. *Journal of Law, Technology, and the Internet*, 4(4): 81–131.

Colson, E. 2019. What AI-driven decision making looks like. *Harvard Business Review*, July. https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. 2017. *Algorithmic decision making and the cost of fairness*. Proceedings of KDD '17, Halifax, NS, Canada. DOI: 10.1145/3097983.3098095.

Cureton, A. 2009. Degrees of fairness and proportional chances. *Utilitas*, 21(2): 217–21.

Duffy, S. H., & Hopkins, J. P. 2013. Sit, stay, drive: The future of autonomous car liability. *Science and Technology Law Review*, 16(3): 453–80.

Foot, P. 1967. The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5: 5–15.

Gertken, J. 2016. Mixed feelings about mixed solutions. *Ethical Theory and Moral Practice*, 19: 59–69.

Gold, N., Colman, A. M., & Pulford, B. D. 2014. Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, 9: 65–76.

Goodall, N. J. 2016. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53: 28–58.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293: 2105–8.

Gustafson, A. 2013. In defense of a utilitarian business ethic. *Business and Society Review*, 118(3): 325–60.

Halstead, J. 2016. The numbers always count. *Ethics*, 126(3): 789–802.

Henning, T. 2015. From choice to chance? Saving people, fairness, and lotteries. *The Philosophical Review*, 124(2): 169–206.

Hill, T. 1992. *Dignity and practical reason in Kant's moral theory*. Ithaca, NY: Cornell University Press.

Hirose, I. 2004. Aggregation and numbers. *Utilitas*, 16(1): 62–79.

Hsieh, N.-H., Strudler, A., & Wasserman, D. 2006. The numbers problem. *Philosophy and Public Affairs*, 34(4): 352–72.

Johnson, D. G. 2015. Technology with no human responsibility? *Journal of Business Ethics*, 127(4): 707–15.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

Kamm, F. M. 1985. Equal treatment and equal chances. *Philosophy and Public Affairs*, 14 (2): 177–94.

Kant, I. 1785/2002. *Groundwork for the metaphysics of morals*. T. K. Abbot (Trans.). New York: Oxford University Press.

Kavka, G. S. 1979. The numbers should count. *Philosophical Studies*, 36(3): 285–94.

Lanteri, A., Chelini, C., & Rizzello, S. 2008. An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83(4): 789–804.

Lawlor, R. 2006. Taurek, numbers, and probability. *Ethical Theory and Moral Practice*, 9: 149–66.

Lazenby, H. 2014. Broome on fairness and lotteries. *Utilitas*, 26(4): 331–45.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. 2019. The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, 160: 377–92.

Lin, P. 2016. Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving: Technical, legal, and social aspects*: 69–85. Berlin: Springer.

Lubbe, W. 2008. Taurek's no worse claim. *Philosophy and Public Affairs*, 36(1): 69–85.

Maitland, I. 2002. The human face of self-interest. *Journal of Business Ethics*, 38(1/2): 3–17.

Martin, K. 2018. Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160(4): 835–50.

Millar, J. 2017. Ethics settings for autonomous vehicles. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence*: 20–34. New York: Oxford University Press.

Miller, R. W. 1998. Cosmopolitan respect and patriotic concern. *Philosophy and Public Affairs*, 27(3): 202–24.

Muller, A. R., Pfarrar, M. D., & Little, L. M. 2014. A theory of collective empathy in corporate philanthropy decisions. *Academy of Management Review*, 39(1): 1–21.

Nyholm, S., & Smids, J. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, 19(5): 1275–89.

Otsuka, M. 2000. Scanlon and the claims of the many versus the one. *Analysis*, 60(3): 288–93.

Parfit, D. 1978. Innumerate ethics. *Philosophy and Public Affairs*, 7(4): 285–301.

Parmar, B. L., & Freeman, R. E. 2016. Ethics and the algorithm. *MIT Sloan Management Review*, 58(1). https://sloanreview.mit.edu/article/ethics-and-the-algorithm/.

Scanlon, T. M. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.

Segall, S. 2016. *Why inequality matters*. Cambridge: Cambridge University Press.

Sidgwick, H. 1907. *The methods of ethics*. 7th ed. London: Macmillan.

Sinnott-Armstrong, W. 2019. Consequentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2019/entries/consequentialism.

Taurek, J. M. 1977. Should the numbers count? *Philosophy and Public Affairs*, 6(4): 293–316.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist*, 59(2): 204–17.

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal*, 94(6): 1395–1415.

Thomson, J. J. 2008. Turning the trolley. *Philosophy and Public Affairs*, 36(4): 359–74.

Timmerman, J. 2004. The individualist lottery: How people count, but not their numbers. *Analysis*, 64(2): 106–12.

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press

Unger, P. 1996. *Living high and letting die: Our illusion of innocence*. Oxford: Oxford University Press.

Villasenor, J. 2014. *Products liability and driverless cars: Issues and guiding principles for legislation*. Washington, DC: Brookings Institution Press.

Williams, B. 1972. *Morality: An introduction to ethics*. New York: Harper and Row.

Williams, B. 1981. *Moral luck*. Cambridge: Cambridge University Press.

. . .

Tobey K. Scharding is an assistant professor at Rutgers Business School, where she has taught business ethics since 2017. She earned her PhD in philosophy at Stanford University. She specializes in ethical decision-making, finance ethics, ethics of risk, and ethics of new technologies. Her book, *This Is Business Ethics*, was published in 2018. A volume coedited with Joanne Ciulla, *Ethical Business Leadership in Troubling Times*, was published in 2019. Her articles have appeared in *Business Ethics Quarterly*, *Journal of Business Ethics*, *Business and Society Review*, and *Rutgers Business Review.*

https://doi.org/10.1017/beq.2020.29 Published online by Cambridge University Press