Sobre as Limitações do Dilema do Bonde para a Avaliação dos Riscos Impostos por Veículos Autônomos

[On the Limitations of the Trolley Problem for Evaluating the Risks Imposed by Autonomous Vehicles]

Renato Rodrigues Kinouchi*

Resumo: Neste trabalho discuto a adequação da versão original do dilema do bonde no que concerne a questão dos potenciais acidentes com veículos autônomos. A principal crítica se refere à impossibilidade de incluir questões relativas aos riscos e incertezas envolvidos em tais acidentes. A seguir, apresento um modelo abstrato que satisfaz os requisitos conceituais da noção padrão de risco, o qual assumirá a forma de um dilema do bonde de natureza probabilística. Mediante a análise de dois exemplos, procuro mostrar que o novo modelo é mais adequado para se avaliar a questão dos riscos impostos por veículos autônomos. Embora não pretenda dar uma solução para o problema colocado, procuro esclarecer como tais veículos pode impor riscos distintos a pedestres e passageiros, uma questão ética não contemplada na versão original do dilema.
Palavras-chave: dilema do Bonde, probabilidade, risco, valores, veículos autônomos.

Abstract: In this paper I discuss the adequacy of the trolley problem original version for the issue of autonomous vehicles accidents. The main criticism concerns the impossibility of including aspects related to risks and uncertainties involved in such accidents. Next, I present an abstract model which satisfies the conceptual requirements of the standard notion of risk and which will take the form of a probabilistic trolley problem. By way of two examples, I defend that this new model is more adequate to evaluate the risks imposed by autonomous vehicles. Although not intending to solve the problem, I try to clarify how such vehicles may impose distinct risks to pedestrians and passengers, an ethical question that the original version of the trolley problem has overlooked.
Keywords: autonomous vehicles, probability, risk, trolley problem, values.

Introdução

Formulado por Philippa Foot e mais tarde modificado por Judith Thomson, o *dilema do bonde* admite várias versões mas a "estrutura básica de todos os dilemas é a mesma: se voce nao agir, cinco pessoas irão morrer; se voce agir, uma pessoa será morta mas

*Professor Adjunto da Universidade Federal do ABC. E-mail: renato.kinouchi@ufabc.edu.br.

ISSN: 2317-9570

DOI: https://doi.org/10.26512/rfmc.v6i2.22105

as cinco serão salvas"(BRUERS e BRAECKMAN, 2014, p. 251)[1]. Originalmente, o dilema do bonde era um dos vários argumentos que Foot (1967) dirigiu à doutrina do *duplo efeito*, usualmente invocada por pensadores católicos para dar sustentação as suas visões sobre o aborto [2]. O duplo efeito em questão refere-se a uma distinção entre os resultados intencionalmente visados por uma ação e os resultados colaterais previstos mas não intencionalmente visados. No que tange a discussão do aborto naquela época, os adeptos de tal noção afirmavam que a cirurgia de histerectomia tinha como efeito previsto, mas não intencional, a morte de eventuais embriões; no entanto, outras intervenções cirúrgicas que levassem ao óbito da criança eram consideradas intencionais, de tal maneira que, nesses casos, os médicos estariam atentando contra vidas inocentes. Convém enfatizar que o dilema do bonde era um dos argumentos mobilizados por Foot em sua discussão da doutrina do duplo efeito, mas, com efeito, havia outros argumentos adicionais, dentre eles a possibilidade de um cirurgião retalhar uma pessoa saudável tendo em vista transplantar seus órgaos para outros cinco pacientes (cf. FOOT, 1967, p. 11). No geral, Foot chega a conclusão de que o *dever negativo* de não fazer o mal prevalece sobre o *dever positivo* de oferecer auxílio; de tal maneira que, para o caso do cirurgião, seria reprovável promover o bem de cinco pessoas à custa da morte de uma outra; mas no caso do bonde, a decisão diz respeito a evitar dois males, e assim desviar a trajetória do bonde seria justificável por minimizar as mortes. Para Foot, a distinção entre deveres positivos e negativos precisava ser levada em conta na discussão do aborto, todavia a autora não se posiciona definitivamente sobre a questão propriamente dita: "Eu não estou argumentando em prol ou contra tais pontos de vista, mas somente tentando discernir algumas das correntes de pensamento que nos movem para frente e para trás. A leviandade dos exemplos nao tem por intenção ofender"(FOOT, 1967, p.18).

[1] Todas as traduções de citações foram realizadas pelo autor deste ensaio.

[2] Segundo a autora: "A doutrina do duplo efeito é baseada na distinção entre aquilo que alguém pode prever como um resultado de sua ação voluntária e aquilo que, em sentido estrito, ele intencionalmente visa (...) As palavras "duplo efeito" se referem aos dois efeitos que uma ação pode produzir: aquilo que é visado e aquilo que é previsto mas de modo algum desejado (...) Diz-se, por exemplo, que a operação de histerectomia envolve a morte de fetos como uma consequência prevista mas não diretamente visada do ato cirúrgico, enquanto outras operações matam a criança e têm por propósito direto eliminar uma vida inocente, uma distinção que tem evocado reações particularmente amargas por parte de pessoas que não são católicas" (Foot, 1967, p. 5-6).

Anos mais tarde, Judith Thomson retomou a discussão iniciada por Foot mas deu novos contornos ao dilema. Para além da configuração original, Thomson (1976, 1985) propôs outros dois casos, a saber: a versão do bonde com *loop* e a versão do *homem gordo* sobre a ponte. Na versão do bonde com *loop*, o trilho lateral onde há uma pessoa isolada se liga novamente ao ramal principal, de tal maneira que o bonde retorna por detrás das cinco pessoas. O resultado disso é a possibilidade de se afirmar que a morte da pessoa isolada, devido ao desvio de rota, poderia ser entendida como um *meio* para se evitar a morte das outras cinco pessoas. Já na versão do *homem gordo*, tal pessoa, cujo tamanho seria suficiente para obstruir o bonde, é empurrada de cima de uma ponte também como um *meio* de evitar a morte das cinco pessoas. Na realidade, deve-se a Thomson as linhas gerais das discussões subsequentes sobre o dilema do bonde (cf. OTSUKA, 2008) e desde a publicação de seus trabalhos uma quantidade enorme de diferentes versões do dilema foram propostos. O principal ponto do debate envolve uma bem documentada inconsistência nas respostas das pessoas entrevistadas quando se comparam a versão padrão do dilema e a versão do homem gordo. Na primeira versão, a maioria dos entrevistados julga permissível a ação de acionar a alavanca que desvia o bonde resultando na morte da pessoa no trilho lateral ao invés das cinco pessoas no trilho principal; entretanto, apenas uma minoria julga permissível a ação de empurrar um homem gordo que se encontra sobre uma ponte acima dos trilhos com o intuito de bloquear a passagem do bonde (HAUSER et al. 2008). Em um trabalho de revisão, Bruers e Braeckman examinaram 35 anos de literatura, classificaram dezessete diferentes grandes versões do problema e mostraram que "a maioria das intuições morais das pessoas não seguem as éticas consequencialistas", um resultado que torna "o dilema do bonde um experimento mental interessante para o estudo das éticas deontológicas" (BRUERS e BRAECKMAN, 2014, p. 252).

Recentemente, o dilema do bonde passou a ser debatido em um contexto muito diferente. Com o desenvolvimento de veículos autônomos, não demorou muito tempo para se estabelecer uma analogia com o experimento moral clássico. Com efeito, suponhamos que um veículo autônomo precise *escolher* entre as opções de atropelar cinco pessoas ou desviar resultando na morte do passageiro: qual deve ser a decisão

a ser tomada? Sem pretender oferecer uma resolução do problema, este ensaio visa analisar a adequação dessa analogia com o dilema do bonde. Na próxima seção, procuro mostrar que o dilema original é incapaz de apreender uma das características essenciais da problemática dos carros autônomos, a saber, seu caráter probabilístico. Isso faz com que a maior parte da literatura produzida sobre o dilema do bonde não seja pertinente ao que realmente se procura determinar na questão dos acidentes com veículos autônomos. Não obstante, acredito que se o dilema do bonde original for transformado em um modelo de caráter probabilístico, então uma pertinente nova questão filosófica aparece, a saber: o que irá definir a distribuição dos riscos de acidentes com veículos autônomos? Essa é a pergunta que tenho a intenção de discutir ao longo deste trabalho.

A analogia entre o dilema do bonde e os veículos autônomos

Em 2015, a prestigiosa revista *Nature* noticiou o desenvolvimento de algoritmos computacionais para o gerenciamento de colisões em situações aparentemente semelhantes ao do dilema do bonde (DENG, 2015). Nesse novo contexto, o dilema original sofre algumas adaptações: o bonde é substituído por um automóvel, a decisão passa a ser feita pelo algoritmo de controle do veículo e as opções oferecidas são entre atropelar cinco pedestres ou desviar em uma manobra fatal para o passageiro. Em 2016, a não menos prestigiosa *Science* divulgou um outro estudo bastante interessante (BONNEFON et al, 2016: cf. SHARIFF et al, 2017), entretanto pouco surpreendente, segundo o qual muito embora os participantes da pesquisa apoiem a tese de que os veículos autônomos devem evitar atropelamentos fatais a pedestres, esses mesmos participantes prefeririam não se locomover em veículos governados por essa diretriz, os chamados *veículos utilitaristas*, preferindo utilizar *veículos autoprotetores*, que não salvariam os pedestres às custas da morte do passageiro:

> Embora as pessoas tendam a concordar que seria melhor para todos que os veículos autônomos fossem utilitaristas (no sentido de minimizar o número de vítimas no trânsito), essas mesmas pessoas têm um incentivo pessoal de se locomoverem em veículos que as protegerão a qualquer custo. Assim, se fossem colocados no mer-

> cado ambos os tipos de veículos, poucas pessoas desejariam utilizar veículos utilitaristas, apesar de preferirem que os outros assim o fizessem (BONNEFON et al, 2016, p. 1575).

Geralmente se reconhece que o dilema do bonde original é um experimento mental muito pouco realista, e por isso o debate também inclui boa dose de ceticismo sobre a adequação de modelos de tipo dilema do bonde para o caso de veículos autônomos. De um ponto de vista jurídico, Casey (2017) afirma que as abstrações filosóficas presentes no dilema do bonde pouco contribuem para os problemas enfrentados pelos engenheiros, sendo mais importante o estudo e o desenvolvimento do aparato legal que irá balizar o uso dessas tecnologias. De um ponto de vista mais filosófico, Nyholm & Smids (2016) assinalam que a analogia carece de sustentação em pelo menos três frentes, a saber: "com respeito às características gerais da situação de decisão, com respeito ao papel da responsabilidade moral e legal, e com respeito à situação epistêmica dos tomadores de decisão" (NYHOLM e SMIDS, 2016, p. 1287). Uma linha de argumentação influente tem sido defendida por Goodall (2014, 2016), segundo a qual a problemática dos veículos autônomos precisa levar em consideração o conhecimento já fartamente produzido na área de análise de risco. Segundo essa abordagem, embora seja razoável supor que a frequência de casos nos quais veículos autônomos terão que enfrentar *escolhas* difíceis – tais como entre matar cinco pedestres ou o ocupante do veículo – será extremamente baixa, por outro lado é praticamente certo que haverá acidentes dos mais variados tipos e com as mais diversas expectativas de mortalidade. Por sinal, as taxas de mortalidade dependem, entre outras coisas, de se o condutor consumiu álcool ou não (duas vezes maiores para condutores alcoolizados), se é homem ou mulher (28% a mais se for mulher), jovem ou idoso (acidentes são quase três vezes mais letais para pessoas acima dos 70 anos em comparação com jovens de 20 anos; cf. EVANS, 2008). Embora os especialistas consigam predizer o resultado dos acidentes mais catastróficos (KOCKELMAN e KWEN, 2002; O´DONNELL e CONNOR, 1996), praticamente tudo o mais que envolve acidentes de trânsito é de natureza probabilística. Ocorre que o dilema do bonde original não serve para tratar de tal questão, dado que o resultado das ações de desviar ou não desviar a trajetória resultam em consequên-

cias letais dadas como certas. A seguir, apresento uma passagem relativamente longa de Nyholm e Smids (2016) que sumariza esse ponto:

> Tão logo começamos a considerar esses vários detalhes adicionais, torna-se claro que aquilo com que estamos lidando aqui não são resultados cujas características são conhecidas com certeza. Na realidade, estamos lidando com situações repletas de incertezas e com numerosas avaliações de risco mais ou menos confiáveis (Cf. GOODALL, 2014, p. 96). Isso significa que precisamos abordar a ética dos carros autônomos usando um tipo de raciocínio moral que não temos razão ou ocasião de usar quando pensamos sobre os casos usuais discutidos na literatura do dilema do bonde. No primeiro caso, precisamos nos engajar em um raciocínio moral sobre riscos e gerenciamento de riscos. Precisamos também de um raciocínio a respeito de decisões sob incerteza. Em contraste, o raciocínio moral empregado por quem lida com um dilema do bonde não é sobre riscos e sobre como responder a diferentes riscos. Nem sobre como realizar decisões que envolvem incerteza. Essa é a diferença categórica entre a ética do dilema do bonde e a ética dos algoritmos e dos acidentes com carros autônomos. Raciocinar sobre riscos e incerteza é categoricamente diferente de raciocinar sobre fatos conhecidos e resultados certos. Os conceitos-chave usados diferem drasticamente no que diz respeito às inferências que eles garantem. E aquilo que selecionamos usando tais conceitos são coisas pertencentes a categorias metafísicas e condições modais diferentes (e.g., risco de dano, de um lado, e dano real, de outro). Portanto, as questões éticas difíceis e específicas que riscos e incertezas engendram não estão em jogo nos dilemas do bonde. Todavia, elas certamente aparecem na ética dos carros autônomos. Permitam-me dar um exemplo. Um número significativo de pessoas pode julgar moralmente inaceitável atropelar um

> pedestre se isso necessariamente matá-lo (Cf. THOMSON, 2008). Não obstante, e se a probabilidade estimada de colisão fatal for de 10%? Ou de apenas 1%? Para muitas pessoas, impor uma probabilidade de morte de 1% a um pedestre inocente tendo em vista salvar cinco ocupantes de um veículo pode parecer uma escolha correta. Os dilemas do bonde não exigem tais juízos. Nos cenários envolvidos nos dilemas do bonde, todos os resultados são assumidos como 100% certos, e desse modo não se precisa refletir sobre a maneira de se confrontar diferentes resultados incertos e/ou arriscados uns contra os outros (NYHOLM e SMIDS, 2016, p. 1286).

Como visto, há razões para considerar o dilema do bonde original como sendo insuficiente para lidar com a problemática dos veículos autônomos. É necessária uma abordagem que seja capaz de estimar os riscos impostos pela utilização de tal tecnologia. Com efeito, abordagem baseada no conceito de risco envolve aspectos valorativos para os quais análise filosófica pode contribuir significativamente, para além dos estreitos limites impostos pelo dilema original (cf. LIN, 2015). A seguir discuto uma nova versão que, acredito, pode lançar alguma luz adicional sobre a questão dos veículos autônomos. Daqui em diante essa nova versão do dilema do bonde original será chamada de *modelo do bonde probabilístico*.

O modelo do bonde probabilístico

Faz-se necessária uma adaptação no dilema do bonde original para transformá-lo em um modelo mais apropriado às situações que realmente serão enfrentadas por veículos autônomos; a saber, é necessário estipular uma probabilidade entre ato de acionar a alavanca e a consequência indesejável do atropelamento das pessoas atadas aos trilhos. Em outras palavras, acionar a alavanca não resultará inescapavelmente em mortes, mas ensejará uma probabilidade para as mortes. Assim, o dilema do bonde original se transforma em um modelo que inclui a noção de risco se considerarmos a relação entre o acionamento da alavanca e a consequência indesejável como sendo de natureza probabilística ao invés de determinista. Para o filósofo sueco Sven Hansson, renomado filósofo

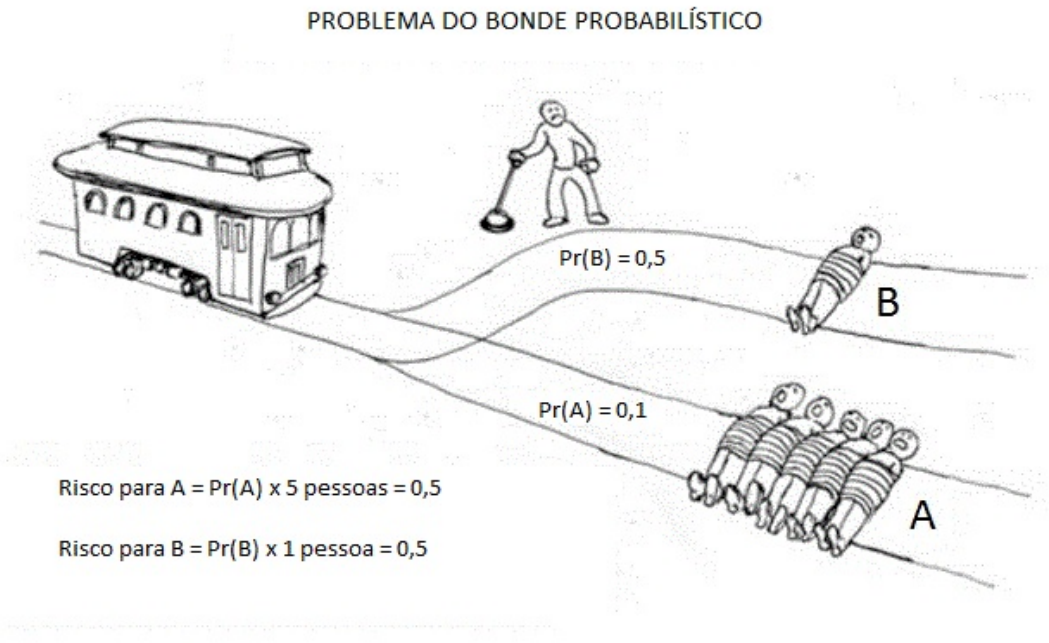
da tecnologia:

> A exclusão da questão de exposição a riscos nas considerações da maior parte da teoria moral pode ser claramente vista nas suposições deterministas comumente feitas nos exemplos usuais de tipo vida-ou-morte empregados para explorar as implicações das teorias morais. No famoso dilema do bonde, você assume saber que se acionar a alavanca então uma pessoa será morta, enquanto que se você não acionar, então cinco pessoas serão mortas (...). Isso claramente contrasta com os dilemas da vida real, onde os problemas das nossas ações que envolvem vidas humanas raramente são acompanhadas por um conhecimento certo sobre as consequências de cursos de ação alternativos (HANSSON, 2012, p. 44).

Segundo a noção padrão usada em estatística, um *risco* é o *valor esperado* de um evento indesejado, o qual pode ser calculado multiplicando-se a probabilidade do evento – expressa por um número dentro do intervalo [0, 1] – por uma estimativa de sua severidade (HANSSON, 2013)[3] . Quando aplicado ao dilema do bonde, o valor esperado do risco de atropelamento consiste na probabilidade de óbito vezes o número de pessoas em cada trilho. Na verdade, a versão proposta é uma generalização da versão original, isto é, o problema do bonde original torna-se o caso especial em que a probabilidade de morte é igual a 1. No dilema original, acionar ou não a alavanca implica que o evento indesejado irá ocorrer inescapavelmente para uma das alternativas. Mas ao trans-

[3] É importante notar que, de acordo com tal definição, a magnitude do risco varia em função tanto da probabilidade do evento quanto da quantidade de dano que o evento pode provocar. Por exemplo, supondo-se que as probabilidades de terremotos no Alasca e na Califórnia fossem iguais, o risco no segundo caso seria muito maior em virtude da grande diferença em termos de perdas materiais e humanas que tal um evento geológico poderia provocar. Cumpre assinalar que o conceito de risco é valorativo (*value-laden*) até mesmo para essa definição técnica, dado que um dos termos da equação deve conter algum tipo de estimativa de um valor (vidas humanas, bens materiais, etc.) sob risco. A respeito desse ponto, Hanson (2004, 2005, 2009, 2012, 2013) costuma enfatizar que "o risco sempre se refere à possibilidade de algo ruim acontecer. Em virtude de ser indesejável, risco envolve valores" (HANSSON, 2013, p. 10). Todavia, ser valorativo não significa deixar de ser factual; em outras palavras, o conceito de risco não se resume exclusivamente a juízos de valor (cf. MÖLLER, 2012). Na verdade, as análises de risco podem ser vistas como socialmente construídas no sentido trivial de que qualquer atividade humana depende de cooperação social, convenções linguísticas, etc., mas, embora seja verdade que as pessoas exibem diferentes percepções de risco acerca de, por exemplo, terremotos, parece estar além de qualquer dúvida razoável que eventos sísmicos de grande magnitude são muito mais frequentes nas áreas de contato entre as placas tectônicas, o que justifica investimentos em projetos de edifícios à prova de terremotos, sistemas de alarmes e outros dispositivos de segurança socialmente adotados em cidades nas proximidades de falhas geológicas.

formar aquela certeza de morte em risco de morte, o valor esperado desse risco pode ser representado pela multiplicação da probabilidade do evento indesejado – e ser atropelado certamente pode ser considerado desagradável – pela estimativa da severidade do acontecimento. Para ilustrar a nova situação, consideremos o enunciado a seguir: "Um bonde trafega sem freios no trilho A e irá atropelar cinco pessoas com probabilidade 10% de que todas elas morram; um transeunte encontra-se ao lado de uma alavanca cujo acionamento desviará o bonde para um trilho B, mas nesse caso o bonde irá atropelar uma pessoa com probabilidade de 50% de que ela morra". A figura abaixo pode facilitar a compreensão do modelo:



Nesta nova versão do dilema, pode-se variar as probabilidades de cada opção, sendo assim também possível criar situações de indiferença de risco onde as opções resultam em um mesmo valor esperado. Para os valores do enunciado acima, os riscos para os ramais resultam em 0,1 x 5 = 0,5 para o trilho A e 0,5 x 1 = 0,5 para o trilho B; ou seja, no tocante ao valor esperado, as opções A e B se equivalem. Em termos técnicos, as opções são indiferentes. O cálculo utilitário, nesse caso, pouco ajuda, não porque falha em fornecer uma estimativa numérica precisa do risco, mas porque o risco é o mesmo para ambas opções. Esse cenário, com efeito, assemelha-se

a um caso comum de escolha de imposição de riscos: por um lado, ao acionar a manivela (opção B) concentramos o risco na pessoa isolada, e, por extrapolação, essa escolha penalizaria minorias em geral; por outro lado, quanta exposição a risco uma maioria deve aceitar (opção A) para proteger indivíduos desafortunados e minorias?

Como já discutido, a analogia entre veículos autônomos e o dilema do bonde original costuma ser considerada insuficiente e pouco proveitosa, todavia o modelo do bonde probabilístico consegue incluir uma questão com implicações filosóficas importantes, a saber, a da distribuição dos riscos impostos pelos veículos autônomos. Cumpre assinalar que embora o bonde probabilístico não seja um modelo estritamente *realista* acerca dos possíveis acidentes, ainda assim inclui, mesmo que de maneira abstrata, um componente de incerteza completamente ausente no dilema original.

Imposição de riscos por veículos autônomos: dois exemplos

A situação descrita no dilema original é certamente pouco verossímil, e quando transposta para a questão dos veículos autônomos costuma tomar a forma de um modelo também pouco verossímil, a saber: um carro autônomo com um passageiro trafega por uma avenida; subitamente cinco pedestres cruzam à frente do veículo; o atropelamento resultará na morte dos cinco pedestres as pode ser evitado se o algoritmo que controla o carro desviar o veículo, que nesse caso colidirá resultando na morte do passageiro. Assim colocada, a analogia entre o dilema do bonde original e os veículos autônomos não costuma receber muito crédito, pois o dilema original acaba restringindo a discussão a situações cujas consequências são dadas como certas.

O problema dos potenciais acidentes com veículos autônomos requer um modelo probabilístico para recobrir as mais variadas situações que implicam em imposição de riscos. Todavia, em se tratando de riscos, também é necessário levar em consideração a severidade deles, isto é, a quanti-

[4]Na realidade, dever-se-ia também considerar aspectos qualitativos – tais como gênero, idade, estado de saúde geral. etc. – o que torna qualquer modelo quantitativo apenas um retrato parcial das situações de risco. No entanto, para os propósitos presentes, confinamos discussão ao aspecto quantitativo.

dade de pessoas afetadas[4]. Com efeito, faz diferença se há apenas um ocupante do veículo e cinco pedestres, ou o caso inverso de cinco ocupantes do veículo e apenas um pedestre. Para ilustrar os riscos impostos por essas variadas condições, consideremos os dois exemplos a seguir.

Exemplo 1: Cinco pedestres subitamente se colocam à frente de veículo autônomo transportando um passageiro. Suponhamos que devido à proteção oferecida pelo veículo ao seu ocupante, o atropelamento não irá causar dano físico ao passageiro. Suponhamos ainda que o acidente não necessariamente causa a morte dos cinco pedestres, mas sim engendra a probabilidade de 50% de que todos eles morram. Nesse caso, o valor esperado do risco consiste na probabilidade de morte (0,5) multiplicada pela severidade do acidente (5 vidas), totalizando 2,5. Por outro lado, se o veículo desviar, no máximo haverá apenas uma morte (a do passageiro), de modo que o valor esperado do risco não pode ultrapassar 1,0; portanto bem menor que em caso de atropelamento.

Se o veículo em questão for autoprotetor, ele nunca se desvia visto que o atropelamento dos pedestres não resulta em ameaça à integridade física do passageiro; em razão disso, o risco de eventuais mortes recai sempre sobre os pedestres. Por outro lado, um veículo utilitarista procuraria minimizar os óbitos, e nas circunstâncias acima descritas sempre desviaria, pois o valor esperado do risco da manobra de desviar (1,0) nunca supera o valor esperado do risco de atropelamento (2,5). Com efeito, é possível que muitas manobras de desviar não resultem na morte do passageiro, mas todas as possíveis mortes virão dessa população. Em resumo, agora o risco recai sobre os ocupantes desses veículos.

Não obstante, pode-se colocar a questão adicional: considerando-se a situação do passageiro, não seria razoável manter a trajetória e "torcer" para que os pedestres saiam vivos do acidente (o que, afinal de contas, pode acontecer com probabilidade de 50%)? Com efeito, há pesquisas que propõem veículos "rawlsianos" (LEBEN, 2017) que optam por desviar exceto em caso de maior ameaça ao passageiro. Um veículo desse tipo não seria exatamente um carro autoprotetor, o qual sempre protege seu ocupante; um carro rawlsiano calcularia os riscos envolvidos para as opções e procuraria evitar uma manobra que causasse a morte do passageiro, mas nos casos em que a manobra significa um risco aceitável ao passa-

geiro, tal veículo desviaria. Todavia, assoma-se uma questão adicional: qual o nível mínimo de risco aceitável que faria um carro rawlsiano desviar? Ao tratar desse problema, Leben (2017) sugere a adoção de um critério de *maximin* (cf. LEBEN, 2017, p. 108-112) segundo o qual, independentemente do valor esperado total do risco aos pedestres (principal critério usado por veículos utilitaristas), um veículo rawlsiano mantém seu curso e protege seu passageiro a partir do ponto em que a probabilidade de morte para o passageiro, em virtude da manobra, for maior que a probabilidade de morte dos pedestres. No entender de Leben: "a diferença conceitual chave é que o carro rawlsiano evita uma alternativa ruim para o sujeito com pior ganho, mesmo se isso resultar em maiores oportunidades para todos os outros" (LEBEN, 2017, p. 112). Todavia, a alternativa proposta por Leben também acarreta problemas dignos de menção. Ao procurar evitar o pior desfecho, um veículo rawlsiano tende a "mirar" veículos/motoristas mais seguros: por exemplo, entre atropelar um motociclista sem capacete (mais inseguro) ou um motociclista com capacete (mais seguro), o carro sempre escolheria a segunda opção, impondo riscos exatamente a quem preza pela segurança.

Exemplo 2: Apenas um pedestre subitamente se coloca a frente de veículo autônomo transportando cinco passageiros; isto é, agora há mais passageiros do que pedestres ameaçados. Suponhamos que devido à proteção oferecida pelo veículo aos seus ocupantes, o atropelamento não irá causar dano físico aos passageiros. Suponhamos ainda que o acidente não irá necessariamente causar a morte do único pedestre, mas sim engendra a probabilidade de 50% de que ele morra.

Nesse caso, o valor esperado do risco consiste na probabilidade de morte do pedestre (0,5) multiplicada pela severidade do acidente (1 vida), totalizando 0,5. Por outro lado, se o veículo desviar, ele pode colocar em risco os cinco passageiros; cabe então indagar: qual probabilidade aceitável para manobra de desviar? Tal como colocado no final da seção anterior deste trabalho: quanto risco uma maioria deve aceitar em prol de uma minoria? O cálculo utilitário leva ao seguinte resultado: quando a probabilidade de morte ao desviar for de 10%, o valor esperado do risco aos passageiros será equivalente ao do pedestre (0,1 x 5 = 0,5). No caso de um veículo utilitarista, o carro deve desviar para valores abaixo de 10% de probabilidade de fatalidade da manobra

para os passageiros; mas para valores acima disso, o veículo deve manter a trajetória. Não obstante, no caso de um veículo rawlsiano, tal como o defendido por Leben (2017), o veículo mantém a trajetória se a probabilidade de morte dos passageiros em virtude da manobra for maior do que a do pedestre (que é de 50%); mas para qualquer valor abaixo de 50%, o veículo desvia, diferindo tanto de veículos utilitaristas – que desviam somente se a fatalidade para os passageiros for menor que 10% – quanto de veículos autoprotetores – que invariavelmente não desviam pois o atropelamento não implica em ameaça à integridade dos passageiros.

As relações de imposição de risco acima descritas, com os valores de probabilidade e número de pessoas utilizados nos dois exemplos, podem ser resumidas da seguinte maneira. Para o exemplo 1, onde há cinco pedestres e um passageiro, veículos utilitaristas sempre desviam e o risco é imposto ao passageiro; já veículos autoprotetores nunca desviam, de modo que o risco é imposto aos pedestres; finalmente, veículos rawlsianos impõe riscos à alternativa cuja probabilidade de mortalidade for menor, o que pode ser interpretado como uma tendência do veículo "mirar" a alternativa mais segura. Para o exemplo 2, onde há apenas um pedestre e cinco passageiros, um veículo autoprotetor novamente impõe risco somente aos pedestres; um veículo utilitarista, por sua vez, desvia se a probabilidade de fatalidade da manobra for menor que 10%, tendendo a impor risco ao pedestre acima desse patamar (entretanto, vale notar que veículos utilitaristas podem ser uma boa alternativa se forem usados prioritariamente para transporte coletivo, tais como ônibus e trens); finalmente, um veículo rawlsiano desvia se probabilidade de fatalidade da manobra para os passageiros for menor que para o pedestre (que é de 50%), e assim impõe risco aos passageiros se a probabilidade de fatalidade da manobra estiver na faixa acima de 10% mas abaixo de 50%.

Tais relações de imposição de risco não podem ser capturadas pela analogia entre veículos autônomos e o dilema do bonde original, pois em tais cenários não há "qualquer incerteza sobre os resultados das decisões, mas uma discussão abrangente sobre algoritmos morais precisará englobar os conceitos de risco, valor esperado e atribuição de culpa" (BONNEFON et al, 2016, p. 1576). Para que tais aspectos sejam levados em consideração é necessário, no mínimo, utilizar alguma adaptação na direção apontada pelo modelo do bonde probabilístico.

Cumpre assinalar que, na verdade, este não se trata de um modelo mais realista, pois ainda recorre a abstrações que não levam em conta os detalhes específicos relativos ao real comportamento de veículos autônomos. Não obstante, por ser de natureza probabilística, ele amplia o escopo da análise e abarca os problemas de imposição de riscos que eram até então negligenciados.

Conclusão

Entusiastas dos veículos autônomos acreditam que sua utilização, em um futuro não muito distante, tornará o trânsito muito mais seguro. Ocorre que tal tecnologia não será perfeitamente segura, em virtude dos inúmeros fatores que contribuem para a ocorrência de acidentes. O dilema do bonde, em sua versão original, coloca em pauta situações extremas, onde qualquer das alternativas disponíveis acarretarão em vítimas fatais. Mas ao focar apenas nessas situações extremas, perde-se de vista que a grande maioria dos futuros acidentes com veículos autônomos, inclusive os acidentes fatais, irão ensejar uma probabilidade de dano à integridade física de passageiros e/ou pedestres. Como visto nos exemplos apresentados, veículos autônomos autoprotetores impõem riscos a pedestres e outros veículos; por sua vez, veículos autônomos utilitaristas podem ser uma boa alternativa para o transporte público, mas para transporte individual tenderiam impor riscos ao passageiro (o que do ponto de vista mercadológico não é nada atraente); finalmente, veículos autônomos rawlsianos (isto é, aqueles que evitam o pior desfecho por meio de algum tipo de cálculo maximin) tendem a "mirar" pedestres ou outros veículos considerados mais seguros – por exemplo, entre colidir com um motociclista usando capacete e outro sem capacete, tal veículo escolheria a primeira opção dado que a probabilidade de morte seria ligeiramente menor.

Tal como alertado no início deste artigo, aqui não se pretendeu resolver esses vários problemas éticos ensejados pela adoção dos veículos autônomos. Mas isso não quer dizer que a filosofia e a ética sejam irrelevantes para o debate, tal como às vezes alguns tecnólogos parecem pensar. Significa dizer, na realidade, que é necessário compreender o novo fenômeno a partir dos conceitos de risco e de incerteza, de modo a conceber modelos mais eficazes para essa tarefa. A proposta do bonde probabilístico é um passo nessa direção. O resultado mais significativo, em meu entender, consiste em se reconhecer que veí-

culos autônomos irão impor riscos de maneira distinta a pedestres e passageiros: uma questão ética para o qual famoso dilema do bonde original é insuficiente.

Agradecimentos: Este trabalho foi escrito durante visita *ao Centre for Philosophy of Natural and Social Sciences* (LSE) com financiamento da FAPESP, bolsa 2017/17081-4. O autor agradece a Roman Frigg por supervisionar a visita e fazer valiosas sugestões à pesquisa.

Referências

BONNEFON, J-F.; SHARIFF, A.; RAHWAN, I. The social dilemma of autonomous vehicles. *Science*, 352 (6293), p. 1573-1576, 2016.

BRUERS, S.; BRAEKCMAN, J. A review and systematization of the Trolley Problem. *Philosophia*, 42, p. 251-269, 2014.

CASEY, B. Amoral machines, or: How roboticists can learn to stop worrying and love the law. *Northwestern University Law Review*, 111, p. 1347, 2017.

DENG, B. The robot's dilemma. *Nature*, 523 (7558), p. 24-25, 2015.

EVANS, L. Death in traffic: Why are the ethical issues ignored? *Studies in Ethics, Law, and Technology*, 2 (1), 2008. Disponível em: https://doi.org/10.2202/1941-6008.1014. Acessado em 30 de agosto de 2018.

FOOT, P. The problem of abortion and the doctrine of the double effect. *Oxford Review, 5*, p. 5-15, 1967.

GOODALL, N. J. Machine ethics and automated vehicles. In: G. Meyer and S. Beiker (Eds), *Road vehicle automation*. Dordrecht: Springer, p. 93-10, 2014.

_____. Away from Trolley Problems and towards risk management. *Applied Artificial Intelligence*, 30 (8), p. 810-821, 2016.

HANSSON, S. O. Philosophical perspectives on risk. *Techné*, 8 (1), p. 10-35, 2004.

_____. The epistemology of technological risk. *Techné*, 9 (2), p. 68-80, 2005.

_____. Technology, prosperity and risk. In: J. K. B. Olsen, S. A. Pedersen and V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology*. Oxford: Wiley-Blackwell, 2009.

_____. A panorama of the philosophy of risk. In: S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (Eds.), *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, p. 27-54. Dordrecht: Springer, 2012.

_____. *The ethics of risk: Ethical analysis in an uncertain world*. New York: Palgrave MacMillan, 2013.

KOCKELMAN, K. M.; KWEN, Y-J. Driver injury severity: An application of ordered probit models. *Accident Analysis & Prevention*, 34 (3), p. 313-321, 2002.

LEBEN, D. A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19 (2), p. 107-115, 2017.

LIN, P. Why ethics matters for autonomous cars. In: J. Markus et al, *Autonomes fahren*, p. 69-85. Berlin/Heidelberg: Springer, 2015.

MÖLLER, N. The concepts of risk and safety. In: S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (Eds.), *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, p. 55-85. Dordrecht: Springer, 2012.

NYHOLM, S. & SMIDS, J. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice*, 19 (5), p. 1275-1289, 2016.

O´DONNELL, C. J. & CONNOR, D. H. Predicting severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, 28 (6), p. 739-753, 1996.

SHARIFF, A., BONNEFON, J-F., RAHWAN, I. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, 1 (10), p. 694-696, 2017.

THOMSON, J. J. Killing, letting die, and the trolley problem. *The Monist, 59*, p. 204-217, 1976.

_____. The trolley problem. *The Yale Law Journal*, 94 , p. 1395-1415, 1985.

_____. Turning the trolley. *Philosophy & Public Affairs*, 36 (4), p. 359-374, 2008