UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS JURÍDICAS E ECONÔMICAS
FACULDADE DE DIREITO

OS DESAFIOS ÉTICOS E JURÍDICOS NO DESENVOLVIMENTO DE VEÍCULOS AUTÔNOMOS

PEDRO MARQUES NEVES

Rio de Janeiro

2021

PEDRO MARQUES NEVES

OS DESAFIOS ÉTICOS E JURÍDICOS NO DESENVOLVIMENTO DE VEÍCULOS AUTÔNOMOS

Monografia apresentado a Universidade Federal do Rio de Janeiro, como requisito para obtenção do Título de bacharel em Direito, sob a orientação do **Prof. Marcelo de Araujo.**

Rio de Janeiro

2021

CIP - Catalogação na Publicação

NP372d

Neves, Pedro Marques
OS DESAFIOS ÉTICOS E JURÍDICOS NO DESENVOLVIMENTO DE VEÍCULOS AUTÔNOMOS / Pedro Marques Neves. -- Rio de Janeiro, 2021.
68 f.

Orientador: Marcelo de Araujo.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Faculdade Nacional de Direito, Bacharel em Direito, 2021.

1. veículos autônomos. 2. inteligência artificial. 3. barreiras éticas. 4. algoritmos de acidente. I. Araujo, Marcelo de, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

PEDRO MARQUES NEVES

OS DESAFIOS ÉTICOS E JURÍDICOS NO DESENVOLVIMENTO DE VEÍCULOS AUTÔNOMOS

Projeto de Monografia, apresentado a Universidade Federal do Rio de Janeiro, como requisito para obtenção do Título de bacharel em Direito, sob a orientação do **Prof. Marcelo de Araujo.**

Data da Aprovação:_____/___/_____.

Banca Examinadora:

_____

Prof. Marcelo de Araujo

_____

Membro da Banca

_____

Membro da Banca

Rio de Janeiro

2021

RESUMO

O presente trabalho tem como proposta analisar os empecilhos éticos e jurídicos enfrentados no desenvolvimento de sistemas de inteligência artificial, no que concerne à tomada de decisão de veículos autônomos. Como é praticamente impossível evitar que veículos, em algum momento, entrem em colisão com outros objetos ou com pessoas ou animais não humanos, fabricantes de veículos autônomos têm de enfrentar o as dificuldades éticas e jurídicas relativas ao desenvolvimento de algoritmos que levem em consideração esse tipo de problema. A pergunta que eles se colocam é sobre como o carro deve reagir na eventualidade de uma colisão. A segurança do motorista deve precedência sobre a seguranças dos passageiros? A segurança do motoristas e passageiros deve ter precedência sobre a segurança de outras pessoas? Será abordado nesta monografia a pergunta sobre até onde essa programação é considerada ética e quais seriam as consequências jurídicas em casos de danos, sejam fatais ou não, uma vez que a legislação atual parece não estar ainda preparada para enfrentar essas questões.

Palavras-chave: veículos autônomos, desafios, inteligência artificial, ética

ABSTRACT

The present work aims to analyze the ethical and legal obstacles faced in the development of artificial intelligence systems, regarding decision making of autonomous vehicles. As it is practically impossible to prevent vehicles from colliding with other objects, people or nun-human animal sat any given time, autonomous vehicle manufacturers have to face the ethical and legal challenges related to the development of algorithms that take this type of problem into account. The question they ask is how the car should react in the event of a collision. Should driver safety take precedence over passenger safety? Should the safety of drivers and passengers take precedence over the safety of others? This work will address the question of how far this programming is considered ethical and what would the legal consequences be in cases of damage, whether fatal or not, since the current legislation does not seem to be prepared to face these issues yet.

Keywords: autonomous vehicles, challenges, artificial intelligence, ethics

LISTA DE FIGURAS

SUMÁRIO

INTRODUÇÃO

Quando pensamos no futuro, uma das ideais que os vislumbramos é a de sistemas de inteligência artificial realizando atividades que as pessoas, de modo geral, não gostariam de realizar, incluindo, por exemplo, carros autônomos, operados por sistemas de inteligência artificial (doravante simplesmente IA). Diversos filósofos e juristas vêm discutindo os empecilhos morais e éticos enfrentados pelos programadores que trabalham com inteligência artificial, no que concerne à tomada de decisão de veículos autônomos. A colisão de um veículo com outro objeto nem sempre pode ser evitada. Por mais que possamos pensar em motoristas dotados de habilidades extraordinárias, existem situações que são inescapáveis como por exemplo a de um pedestre que atravessa fora da faixa ou a de um animal que invade a pista para veículos. Os reflexos e instintos do condutor fazem com que ele, em uma tentativa de evitar a colisão, tome uma ação de último segundo, guinando o veículo para uma direção, o que pode acarretar em um acidente diverso, como por exemplo bater em um muro ou atropelar outra pessoa.

Nos casos dos carros autônomos (doravante AV, de *autonomous vehicles*) entretanto, esperamos que todas as decisões tomadas por meio de inteligência artificial sejam as que evitem acidentes e situações críticas. Entretanto, podemos descartar – ao menos por enquanto – que o sistema de IA do veículo é capaz de escolher a melhor decisão com o grau de desenvolvimento tecnológico que nossa sociedade possui à disposição, afinal, se a máquina é treinada e alimentada com dados feitos por humanos, poderia ela realmente tomar decisões por conta própria que não resultem em nenhum tipo de fatalidade?

Como ainda não alcançamos tal grau de proeza tecnológica que ateste a devida segurança para que um AV consiga pilotar completamente livre de interferência humana sem apresentar riscos, devemos, por enquanto, trabalhar com a hipótese mais provável em um futuro próximo, qual seja AVs sendo “pilotados” por um humano como condutor, que interviria em situações nas quais a IA do veículo não consiga exercer sua função sem apresentar riscos de acidente.

Portanto, uma vez que o veículo será guiado por códigos programados que buscam ler o ambiente para realizar o ato da condução, podemos imaginar situações inesperadas em que os programadores não saberiam como o sistema reagiria, exigindo da IA uma tomada de decisão que esbarra no campo da moralidade. Tal situação compreende um dilema clássico, atribuído à filósofa Philippa Foot[1], no qual temos um trem desgovernado em alta velocidade prestes a atingir um grupo de cinco pessoas. Todavia, o condutor, ao perceber a situação, pode puxar uma alavanca, que desviaria o sentido do trem para outra direção, matando apenas uma pessoa, e consequentemente salvando dos outros, que não serão atropelados pelo trem. O que seria, então, correto fazer?

Transpondo esse dilema para uma situação fictícia envolvendo um carro autônomo, nos deparamos com uma situação semelhante, que exige a mesma complexidade moral. Por exemplo, imaginemos que em uma rua um AV irá colidir com uma senhora, independente do sistema de frenagem atingir seu máximo potencial. À sua direita, temos duas crianças saindo da escola e à esquerda, um outro carro vem em velocidade igual. Entretanto, colocamos aqui um fator a mais na balança: a vida do condutor. Afinal, quem iria pilotar um carro que escolheria guinar à esquerda, buscando uma colisão frontal, ceifando a vida do condutor para salvar outra?

Esforços estão sendo feitos para desenvolver algoritmos que trabalhem em situações parecidas com o dilema do bonde (DENG, 2015). Diversos especialistas internacionais reconhecem que um amplo debate acerca desse dilema ético é necessário. Para eles, existe uma probabilidade de que os carros acabem por enfrentar os *no-win scenarios*, situações nas quais haverá um prejuízo que independe da decisão do veículo, mas seria desarrazoado pensar que os empecilhos que advém da ciência da computação vão se solucionar sozinhos sem uma discussão aberta sobre ética e qual escolha por parte do carro seria a mais correta.

Segundo Patrick Lin:

> Se você reclama que carros robôs provavelmente nunca estariam no dilema do bonde – que as chances de se ter que fazer tal decisão seja minúscula e valem a pena serem discutidas – então você não está vendo a questão com clareza. Programadores ainda precisam instruir um carro autônomo em como agir na totalidade de toda a gama de cenários possíveis, além de estabelecer princípios norteadores para cenários não previstos. Então os programadores terão que eventualmente confrontar essa decisão, mesmo que nós motoristas humanos nunca tenhamos que fazer isso no mundo real. (tradução nossa) [2]

[1] FOOT, Philippa. *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices*. England: Oxford University Press. 1967, p. 4.
[2] LIN, Patrick. *The Ethics of Autonomous Cars*. The Atlantic, 2013, p. 5.

Noah J. Goodall (2014) defende que os dilemas relacionados aos AVs precisam levar em consideração os conhecimentos da área de análise de risco. É possível que a frequência dos casos que os carros enfrentarão escolhas morais seja muito remota, mas em contrapartida, muitos acidentes, no longo prazo, com certeza irão ocorrer, o que evidencia a necessidade do debate. Ainda que essas situações tenham chance baixas de ocorrer, com milhões de carros autônomos nas ruas, eventos com baixa probabilidade podem acabar se concretizando (BONNEFON *et al*, 2016).

Como, então, desenvolver tais carros sabendo que caso uma situação em que não haja saída aconteça, a solução de tirar a vida de um indivíduo foi pensada por alguém? Em um outro interim, temos também que vislumbrar de que forma a legislação irá preencher as lacunas deixadas por essa nova tecnologia. Afinal, em um caso de acidente, de que forma será tratada a responsabilidade civil?

Com essas indagações, esta monografia visa apresentar as barreiras enfrentadas no desenvolvimento de carros autônomos em face das mais diversas situações que podem ocorrer, sendo necessária uma tomada de decisão na "área cinza" do campo da ética, além de suas implicações jurídicas

1. HISTÓRIA, CONCEITO E COMO FUNCIONAM OS VÉICULOS AUTÔNOMOS

1.1 Breve História dos Veículos Automotores até sua Automação

Antes de entrar no tema que dá a essência a este trabalho, faz-se necessário prestar um contexto histórico, além de mostrar de que forma os AVs são programados para tomar suas decisões. O automóvel em si tem contexto histórico datado do século XVIII, mais precisamente no ano de 1769, com a criação do motor a vapor, criado por Nicolas-Joseph Cugnot. Diversas variações de motor com diferentes fontes de energia foram sendo usadas até 1882, quando o italiano Enrico Bernardi criou o primeiro motor a base de petróleo.

Três anos depois, em 1885, Karl Benz desenvolveu o primeiro veículo que funcionava a base de petróleo, que é considerado o primeiro. Não foi até 1908, entretanto, que o primeiro veículo de produção em massa, o *Ford Motel T*, passou a ser fabricado.

Somente em 1939, quase duzentos anos após a criação do primeiro motor, que a ideia de veículo autônomo ganhou fama através do mundo, na *World Fair*, um evento internacional designado para exibir diversas realizações humanas. Lá, a *General Motors* exibiu veículos elétricos movidos por campos eletromagnéticos gerados por circuitos que ficavam no trajeto. A partir de então, a evolução dos carros parece tender a ideia apresentada em 1939, qual seja, o veículo cada vez mais automatizado, facilitando o trabalho do condutor (FINN, 2010).

Com a passagem das décadas, os veículos acabaram por ganhar diversas funcionalidades que corroboraram para uma maior segurança do seu passageiro e do pedestre, visando sempre, como concepção final, a automação total, como, por exemplo, o freio de emergência automático, avisos de colisão, assistência de permanência de faixa, dentre outros (OZGUNER et al, 2011). Para Hannah YeeFen Lim (2018), a ideia central dos AVs não é ser uma opção mais segura do que os condutores humanos, mas apenas a conveniência de não termos mais, como sociedade, que praticar o ato de dirigir. Isso se comprova observando que uma grande parte dos mecanismos de segurança que são implementados nesses veículos, como os citados acima, já se encontram nos veículos convencionais.

Atualmente, temos diversas multinacionais investindo no setor de automação de veículos, sendo usadas, como exemplo, as empresas Google, Toyota, Tesla, Audi, Uber,

Hyundai, Mercedes, dentre outras. Imperioso ressaltar que, das citadas, a Tesla e a Uber já figuraram em acidente de AV, tendo o seu carro fabricado como vetor ativo do acidente, e resultando, inclusive, em vítimas fatais. Por conta desses testes, muitos especialistas defendem que esses veículos, mesmo em um futuro próximo, só poderão transitar em condições específicas, como: vias específicas para trânsito de veículo automotor, velocidade máxima limitada e condições meteorológicas padrão.

1.2 Conceito de Veículo Autônomo

Um AV é um automóvel capaz de, através de sensores, detectar o espaço físico ao seu redor para, assim, operar sem o controle direto de um condutor humano. Não é necessário em momento algum que um passageiro humano assuma o controle do veículo, nem que alguém esteja sequer presente no automóvel. Dessa forma, um carro autônomo pode ir a qualquer lugar tradicional por si, fazendo tudo o que um motorista humano com um certo grau de experiência consegue fazer.

Existe uma distinção que deve ser feita entre os veículos autônomos e os semiautônomos, que o *U.S. Department of Transportation* (Departamento de Transporte dos Estados Unidos) e a *National Highway Traffic Safety Administration* (Administração Nacional de Segurança do Tráfego), ambas Norte Americanas, fazem. Essa separação se dá com base em uma classificação criada pela *Society of Automotive Engineers* (Sociedade de Engenheiros Automotivos*),* que separa os veículos em seis níveis, do 0 ao 5, com base na capacidade do automóvel, sendo eles:

- Nível 0: Sem nenhum tipo de automação. É o veículo comum, em que o condutor tem o controle total do veículo, tendo toda a responsabilidade inerentes à atividade de dirigir. A grande maioria dos carros atualmente se enquadram nesse nível.

- Nível 1: Assistência de direção. O veículo é controlado tem a maioria das atividades exercidas pelo condutor, porém, existem algumas tecnologias disponíveis no automóvel que prestam algum tipo de assistência. Pode ser usado como exemplo os automóveis de hoje em dia que possuem alguma função de auxílio embutida, como, por exemplo, o *Cruise*

Control[3]. Apesar de poder passar o controle da direção para o veículo em algumas situações, o condutor continua completamente responsável pelo que vier a acontecer, devendo ficar atento ao seus arredores e preparado para controlar o automóvel nos casos em que o sistema puder vir a falhar.

- Nível 2. Automação Parcial. O veículo possui funções automatizadas combinadas, como, por exemplo, ter a sua direção e velocidade controladas por um ou mais sistemas de assistência ao motorista, mas um ser humano controla os outros elementos da direção. Todavia, o condutor deve permanecer envolvido com a tarefa de dirigir e monitorar o ambiente o tempo todo. Exemplos disso são os sistemas de frenagem de emergência e sistema de manutenção de faixa. O condutor ainda é responsável pelo controle e operação segura do veículo e deve ser capaz de intervir quando necessário.

- Nível 3: Automação Condicionada. Os veículos de nível 3 e acima são considerados, *lato sensu*, automatizados. Para obter essa classificação, os veículos desse nível precisam ser capazes de monitorar o ambiente de direção ao seu redor. O motorista não é obrigado a monitorar o ambiente externo, mas é necessário estar pronto para assumir o controle do veículo a qualquer momento e dar instruções, para fins de segurança. Um automóvel desse nível não é capaz de tomar uma decisão sozinho, portanto, para uma ultrapassagem, por exemplo, o carro detectaria um outro veículo a frente e o condutor tomaria a decisão de desacelerar ou ultrapassar.

- Nível 4: Alta Automação. O veículo é capaz de realizar todas as funções de direção sob certas condições. O veículo irá parar se os sistemas falharem. O motorista tem a opção de controlar o veículo, especialmente quando as condições confundem o sistema do veículo, como obras ou desvios de estradas, por exemplo.

- Nível 5: Automação Total. Nesse caso, o último estágio da autonomia, ainda não alcançado, em que o veículo é capaz de realizar todas as funções de direção sob todas as condições, podendo operar em completa segurança sem a necessidade de nenhum controle humano. O condutor teria a opção de controlar o veículo, mas não se espera em situações estritamente

[3] O *Cruise Control* é um sistema que mantém a velocidade de condução de um veículo previamente programada. Uma vez atingida e memorizada a velocidade pretendida, pode-se retirar o pé do acelerador, permitindo assim um maior conforto da condução em estrada ou em viagem.

emergenciais, não disponível a qualquer momento durante o trajeto. Os veículos de nível 5 tem automação total e, por isso, não possuem pedais, volantes, ou controles para que um humano assuma o comando. Os sistemas controlados por computador tomam as próprias decisões, com nenhuma intervenção humana. O motorista, neste caso, atua como um passageiro, ele não tem que observar e monitorar o trânsito, precisando apenas informar o destino no sistema. Isso vale tanto para veículos ocupados quanto desocupados, significando que o motorista pode solicitar o carro sob demanda por seu telefone para o local desejado. este caso, o motorista não precisa estar a bordo, e um veículo pode dirigir sozinho para qualquer destino legal e tomar sua própria decisão ao longo do trajeto (HUCKO, 2017).

1.3 Funcionamento dos Veículos Autônomos

Um AV nada mais é do que a soma de diversos componentes tecnológicos, muitos desses, inclusive, já disponíveis no mercado, que, utilizados em conjunto com a inteligência artificial e os dados por ela armazenados possibilitam a transição do automóvel por um caminho. A importância de entender suas funcionalidades se dá na medida em que, quanto mais conhecemos o que faz o veículo funcionar, melhor podemos regular essa nova tecnologia, pois, a partir do momento em que entendemos o que cada componente faz, conseguimos começar a compreender seus limites, e assim melhor definir parâmetros de segurança, bem como legislar.

Entretanto, mesmo com todos os componentes à disposição, frisa-se é quase impossível prever como a IA irá reagir diante das inúmeras situações que o sistema pode se deparar, vez que tais situações beiram o infinito. Portanto os dispositivos que serão abordados nesse capítulo servem de alicerce para que os avanços tecnológicos de programação e *machine learning* girem em torno deles.

Então, quais são e como os componentes tecnológicos do carro irão fazer o seu papel de guiá-lo? Tudo ocorre em um procedimento denominado *sense-plan-act* no qual o veículo, antes de tomar a ação necessária para seu correto prosseguimento, se utilizará de seus sensores que colherão dados externos para serem analisados pelos algoritmos do software e, por fim, formular uma decisão, que abrange desde a frenagem até uma guinada de direção. O *sense-plan-act* é um processo contínuo, que ocorre incontáveis vezes em uma fração de segundos, pois o ato de sensoriamento deve sempre estar checando por novas informações para alimentar

os sistemas do veículo. Para Hannah YeeFen Lim (2018), o principal desafio desse sistema será se adaptar a um ambiente humano de veículos, pois há uma miríade de variáveis acrescentáveis, tais como pedestres, ciclistas e outros veículos conduzidos por humanos, todos esses fatores em velocidades distintas. Neste processo também serão levados em conta condições meteorológicas, obras na rodovia (trabalhadores, placas de sinalização, cones ou até mesmo buracos) e acidentes.

1.3.1 O Sensoriamento

O grande desafio do veículo, portanto, será distinguir os diversos fatores mencionados, para que o carro saiba planejar o que irá encontrar pela frente. A responsabilidade deste ato está incumbida ao sensoriamento, que alimenta o sistema do veículo com as informações que coleta pelos componentes tecnológicos que preferencialmente devem estar funcionando em conjunto, embora, pelo custo atual, não é uma alternativa economicamente viável para o consumidor médio.

- Sistema de Câmeras: o sistema de câmeras é responsável por coletar as imagens que serão utilizadas pelo software para processar o que está acontecendo ao redor do veículo, visando orientar sua decisão. O veículo preferencialmente deve ter diversas câmeras de 360º para que não haja falta de informação. Além de sua função de identificar objetos, as câmeras podem prever suas trajetórias imediatas quando usam scripts e algoritmos avançados (MUJICA, 2014).

- Sistemas LIDAR (*light detection and ranging*): é um tipo de sistema que determina a distância do objeto através da emissão de raios lasers e o tempo que esses raios demoram para ser refletidos pelos objetos que estão ao redor do veículo. Podemos observar esse princípio também na natureza. Os morcegos enviam ondas sonoras e são capazes de calculando a distância dos objetos de acordo com a rapidez com que a onda volta para eles. Se sabemos a velocidade com que o som ou as ondas de luz viajam em um ambiente específico, podemos localizar exatamente a que distância está o objeto que a onda atingiu (HUCKO, 2017). De acordo com Lim, é o sistema mais importante, por ser extremamente preciso, já que consegue rastrear objetos enquanto eles se locomovem, permitindo um melhor mapeamento da área ao redor e, portanto, uma previsão correta do local que os obstáculos estão se direcionando. Uma outra vantagem em relação as câmeras também é

sua maior precisão em um cenário de escuridão. Entretanto, as desvantagens incluem um menor alcance de visão e dependência da reflexividade dos objetos.

- Sistemas de Radar: funcionam através de emissão de ondas de rádio para detectar objetos ao redor do veículo. Apesar de não serem tão precisos como o LIDAR, o radar possui um maior alcance, podendo servir como uma forma de pré-detecção de objetos que serão tidos como obstáculos em momento próximo (REINA *et al*, 2015).

- Sistemas Infravermelhos: componente vital para detectar seres vivos durante o trajeto do veículo, o que seria ideal para complementar os outros tipos de sistemas, devendo serem usados em conjunto.

- Mapas digitais em alta definição: mapas utilizados como principal guia para o veículo. Por conta do nível de detalhamento desses mapas, o veículo consegue saber o caminho que tem que percorrer de forma "geral", devendo ser guiado pelos outros sistemas acima mencionados para evitar colisões com outros automóveis. Os mapas, dessa forma, devem ser frequentemente atualizados, para que o veículo não acabe por cometer algum erro de direção. Uma grande vantagem que os mapas podem trazer é a possibilidade de que o veículo planeje seu caminho com maior precisão, como, por exemplo, saber se existe um sinal ou uma lombada em determinada parte do trajeto, podendo já preteritamente se preparar para diminuir sua velocidade e se adequar às circunstâncias.

Os pontos acima listados correspondem aos principais dispositivos evidenciados que são utilizados para guiar o trajeto dos AVs. É na fase de sensoriamento também ocorre um fenômeno que conhecemos como *Big Data*.

Conforme explicado neste subcapítulo, os AVs baseiam-se inteiramente nos dados que recebem por meio das tecnologias listadas acima para que uma decisão seja tomada em relação ao seu percurso. Portanto, o *Big Data* é o que ajuda o veículo a fazer o uso de seus sensores.

Sem acesso a um fluxo constante de *Big Data*, os veículos seriam inúteis na estrada, pois não saberia o que fazer com os dados que recebem. Portanto, a transmissão de dados deve

ser constante, inclusive entre os próprios veículos, que depositam as informações coletadas em uma nuvem para compartilhá-las, no intuito de que outros sistemas tenham acesso para basear sua tomada de decisão.

Usando os dados coletados, um AV pode construir estratégias para muitas situações possíveis na estrada. Como alguns exemplos, pode ajudar evitando a formação de engarrafamentos, compartilhando condições climáticas ou informando à respeito de possíveis emergências na estrada.

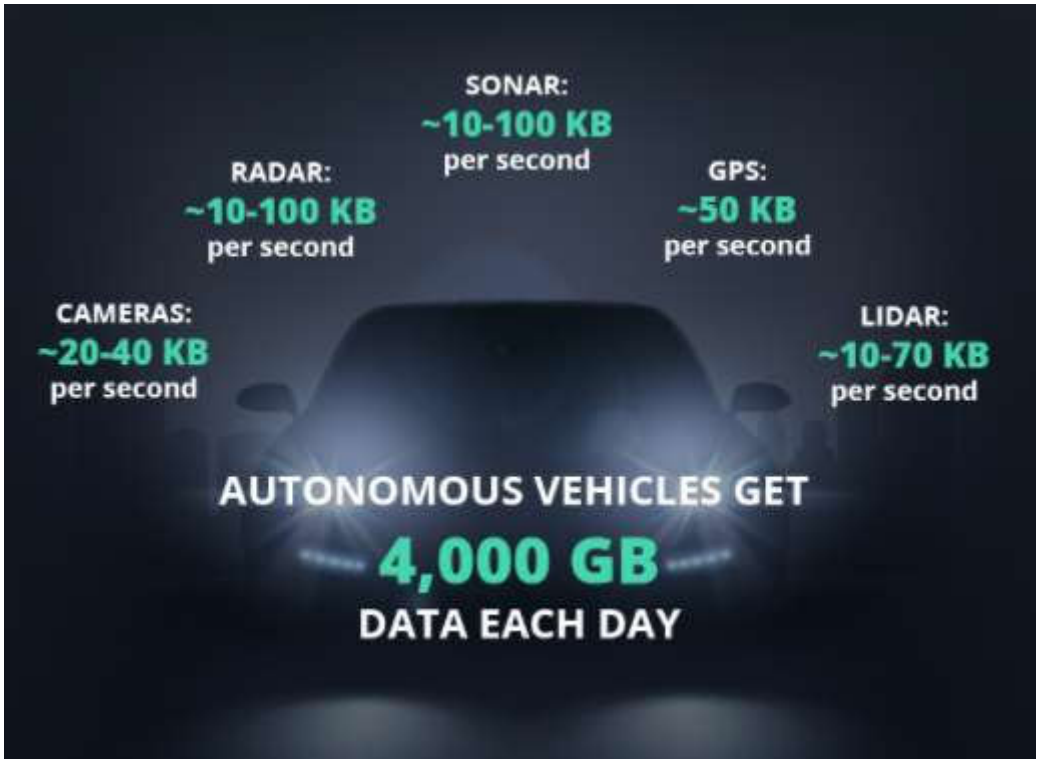

Figura 1 – A quantidade de informação utilizada e repassada por veículos autônomos. Cerca de 4 mil *gigabytes* são transferidos diariamente. Fonte: https://www.intellias.com/how-big-data-in-autonomous-vehicles-defines-the-future/

Existem, entretanto, algumas preocupações por parte da sociedade científica quanto ao uso do *Big Data*. Por precisarem adquirir uma grande quantidade de dados sobre o comportamento de outros veículos, é possível que, acidentalmente, o sistema consiga obter dados sobre outros usuários na estrada.

Fillipo Sio (2016) argumenta que o *Big Data* é fundamental para garantir a segurança e responsabilização do AV através da aquisição e armazenamento dos dados. Entretanto, Sio

destaca que os sistemas podem acabar por acidentalmente adquirirem muitos dados sobre outros usuários na estrada ao realizar essa interação de troca de informações. Dessa maneira, alguns problemas podem surgir. Em suas palavras:

> Dois riscos éticos e sociais destacados na literatura ética sobre privacidade e proteção de dados (por exemplo, van Den Hoven 2008) estão claramente presentes também nos casos dos sistemas de condução automatizados. Em primeiro lugar, o sistema de condução automatizado pode ser alvo de ataques cibernéticos ou hackers. Em segundo lugar, a aquisição e armazenamento em massa de dados pessoais dos usuários podem ameaçar suas autonomias morais de duas maneiras: a) criando uma assimetria de informação: uma grande quantidade de informação sobre pessoas individuais pode ficar à disposição de quem possui ou controla infraestruturas de transporte. Estas informações podem ser utilizadas em benefício dos cidadãos, mas também existe o risco de serem recolhidas e utilizadas contra os interesses das minorias ou mesmo da maioria das pessoas e na violação de seus direitos; b) criando um desequilíbrio de poder: da mesma forma, um aumento dramático nas capacidades de informação de governos ou outras agências pode aumentar sua capacidade de promover a segurança e o bem-estar do cidadão, mas essa capacidade também pode ser usada para controlar, coagir, explorar , discriminar e até oprimir as pessoas (tradução nossa).[4]

Podemos perceber, portanto, que diversas questões éticas de privacidade e proteção de dados podem emergir dessa fase de sensoriamento, constituindo um desafio para os legisladores estabelecer os limites nas trocas informacionais feitas pelos softwares da IA dos veículos.

1.3.2 A Fase de Planejamento

Passada a fase do sensoriamento, o veículo pode planejar e decidir sua próxima ação com base nas informações coletadas pelos sensores. Essa é, portanto, a fase de planejamento, quando o automóvel se utiliza do que foi coletado para se situar no ambiente escaneado, devendo, ainda, levar o fator movimentação, como um carro vindo na direção contrária. Através

---

[4] SIO, Fillipo Santoni de. *Ethics and Self-driving Cars*: A White Paper on Responsible Innovation in Automated Driving Systems. 2016. Texto original: "T*wo ethical and societal risks highlighted in the ethical literature on privacy and data protection (e.g. van Den Hoven 2008) are clearly present also in the case of ADS (automated driving systems). Firstly, ADS may be the target of cyber attacks or hacking. Secondly, the massive acquisition and storage of personal data about road users may threaten their moral autonomy in two ways: a) by creating an information asymmetry: a huge quantity of information about individual persons may become available to those who own or control transport infrastructures. This information may be used to benefit citizens, but there is also a risk that it will be collected and used against the interests of minorities or even the majority of people and in violation of their rights; b) by creating an imbalance of power: similarly, a dramatic increase in the information capabilities of governments or other agencies may enhance their capacity to promote the citizen's safety and well-being, but this capacity may also be used to control, coerce, exploit, discriminate and even oppress people."*

de milhares de cálculos por segundo, a inteligência artificial do veículo analisa as opções e assim toma uma decisão.

Lim defende a utilização de um processo conhecido como *machine learning*, podendo ele ser extremamente útil para a tomada de decisão. O *machine learning*, que será abordado no decorrer desse trabalho, nada mais é do que a própria inteligência artificial, ao analisar seus dados, verifique que exista padrões de comportamento em suas tomadas de decisão e, sendo essa decisão correta, começa a criar uma tendência de comportamento voltada para aquela ação que obteve maior sucesso. Conforme uma situação que exija um novo comportamento do veículo surja, deve a inteligência se adaptar, saindo, de certa forma, de sua "zona de conforto", e aprendendo como se portar em diferentes situações.

O principal problema nesse caso é que as companhias que desenvolvem essas tecnologias não são transparentes em relação aos seus algoritmos ou até sua própria inteligência artificial. Em outras palavras, por conta da constante necessidade de criar a melhor tecnologia e competir com outras concorrentes, as empresas não revelam abertamente como fazem para que a inteligência artificial do veículo tome determinada decisão, pois não querem que outras fabricantes possam se utilizar de alguma forma de seu produto. Segundo Lim, tal fato constitui um problema do ponto de vista social e legal, pois os usuários dos veículos não possuem conhecimento do que estão pilotando, gerando dúvidas referentes à segurança nas rodovias.

É nessa fase que o problema ético proposto nessa monografia ocorre, uma vez que o planejamento consiste na tomada de decisão da IA com as informações que o veículo conseguiu reunir.

1.3.3 A Fase de Ação

Por fim, temos a fase de ação, que é autoexplicativa. É a etapa em que o veículo irá executar aquilo que foi calculado e programado seguindo todos os dados coletados durante a fração de segundos em que as fases passadas se perpetuaram.

2. MORALIDADE E SUAS TEORIAS QUE EMBASAM A PERCEPÇÃO HUMANA NA TOMADA DE DECISÃO DOS VEÍCULOS AUTÔNOMOS

O segundo capítulo visa, de forma sintetizada, explicar a conceituação da moralidade, bem como ramificar as teorias que derivam desses conceitos, e que embasam a forma projetar os algoritmos de um AV frente a uma situação de colisão. Imaginemos, portanto, a situação central que esse trabalho pretende analisar, qual seja, um iminente acidente que, independentemente da decisão tomada pelo carro, acarretará em algum prejuízo mortal: seria correto tomar uma decisão baseada no utilitarismo ou na deontologia? Antes de analisar tal conjuntura, todavia, devemos primeiro ponderar acerca do conceito da palavra moralidade, para posteriormente entendermos o que significa agir de acordo com tal princípio.

2.1 Algumas Posições em Filosofia Moral

Tentar explicar o significado de moralidade é um dos grandes desafios da filosofia e sociologia, pois tal conceituação é cercada de debates que perduram há anos. Afinal, tal palavra engloba uma alta carga subjetiva, pois é, em sua essência, abstrata. Existem debates, inclusive, sustentando que a palavra ética tem o mesmo significado que a moral, e, em contrapartida, diversos filósofos também dedicam obras inteiras para exemplificar a divergência que existe entre elas.

Kant (1994) defende que a moralidade nos obriga incondicionalmente, independentemente de sentimentos, interesses, ou aspectos de nossa psicologia. A “lei moral”, para Kant, é uma norma universal, válida para todos os seres racionais.

Para Tugendhat (1998), a moralidade é intrínseca as exigências recíprocas da vida em sociedade. O comportamento moral não é baseado em um conjunto de regras de ação, mas em um sistema de exigências em que todos os indivíduos desse grupo considerem essa regra como justificada. Portanto, o conceito de moral aqui deve ser compartilhado por todos os membros da sociedade (SILVEIRA, 2012).

Já para Durkheim, a moralidade é uma regra de conduta pré-estabelecida que se manifesta na consciência humana, sem que seu pensador se questione da razão para tal ou se

deve fazer ou não tal atitude, pois aquilo, para ele, é o certo. Ao contrário de Kant, Durkheim defende que o dever não é a única alavanca da moral, mas também o desejo (WEISS, 2007).

Como bem preceitua Rachels (2003), definir o que é moralidade de forma simples e direta é impraticável, já que existe uma miríade de teorias divergentes, cada qual com sua própria visão do que é ou não. Portanto, a concepção de um doutrinador acerca da moralidade está fadada a ofender outra teoria. Com isso, ele elabora o que podemos chamar de "concepção mínima de moralidade", o que, nas suas palavras, se traduz:

> (...) moralidade é, minimamente, o esforço em guiar a conduta do indivíduo por meio da razão – ou seja, fazer algo para o qual haja as melhores razões para fazê-lo – enquanto ao mesmo tempo se dá um peso igual aos interesses de cada indivíduo que será afetado pelo que alguém faça. Isto nos dá, entre outras coisas, um retrato do que significa ser um agente moral consciente.[5]

No mesmo interim, Gert (2020), define que moralidade é (ou seria) o código de comportamento que atende à seguinte condição: todas as pessoas racionais, sob certas condições especificadas, o endossariam. O que, segundo ele, é um esquema básico plausível para definições de "moralidade" no sentido normativo.

Portanto, para fins de praticidade, utilizaremos aqui as duas últimas concepções tratadas, por Rachels e Gert, que, em termos gerais, definem bem o que é uma ação moral do ponto de vista social. Quanto a diferenciação de ética e moral, conforme já introduzido, Canto-Sperber (2002), é um defensor de que existe sinonímia entre elas. Entretanto, há uma espécie convenção adotada por autores para diferenciar os seus sentidos.

Para esses autores, moral seria o fenômeno social, já ética seria a reflexão filosófica acerca da moralidade. As comunidades humanas são regidas por um conjunto de regras de conduta e proibições, que devem ser seguidas, com base na moral. Entretanto, pode-se procurar entender como os homens legitimaram esses padrões de comportamento, quais condições sociais que os tornaram possível ou necessário, e, a essa reflexão filosófica e científica, se da o nome de ética (LA TAILLE, 2007).

---

[5] RACHELS. *Os Elementos da Filosofia da Moral*. Barueri, SP: Manole, 2006, p. 15.

A pergunta que permanece, todavia, é: seria esse conceito de moralidade universal? Podemos então abranger todas as sociedades e diferentes culturas como portadoras do mesmo padrão de código moral?

Lawrence Kolbergh, em 1958, chegou a constatar em seus estudos diversos padrões morais presentes nos seres humanos que independiam de divergências culturais, o que o levou a defender uma universalidade do caráter moral, atribuindo exclusivamente a moralidade ao sentimento de justiça, e nada mais (ARAUJO, 2000). A crença de que "o senso moral rudimentar é universal e surge cedo" (BLOOM, 2010, p. 490) é suportada por diversas evidências de diferentes campos de estudo. Em suma, alguns autores defendem que os precursores da moralidade são inatos e até mesmo "as crianças interpretam o mundo de uma maneira pró-moral" (PREMACK, 2007, p. 164).

Tal fato é inato pois, segundo Premack, até mesmo os recém-nascidos, em seus primeiros meses de vida, demonstram atitudes pró moral, sem ter absolutamente nenhum tipo de experiência social para aprender as normas do comportamento social, isto é, virarem seres morais através da vivência com outros humanos.

Entretanto, muitos pensadores discordam dessas observações, defendendo que diferentes culturas possuem códigos morais diversos, rompendo a tal ideia da moral universal. Para eles, o código moral está atrelado à cultura, e, por esse motivo, recebeu o nome de "Relativismo Cultural". Rachels (2003) distingue seis argumentos feitos por autores que defendem essa teoria:

- Diferentes sociedades possuem diferentes códigos morais.

- O código moral de uma sociedade determina o que está certo dentro daquela sociedade, ou seja, se o código moral de uma sociedade diz que determinada ação está certa, então a ação está certa, pelo menos naquela sociedade.

- Não existe um padrão objetivo que pode ser empregado para julgar o código de uma sociedade melhor do que outros.

- O código moral de nossa própria sociedade não possui um status especial; é somente um entre muitos.

- Não existe uma "verdade universal" na ética, ou seja, não há verdades morais que são tomadas por todas as pessoas em todos os tempos.

- É mera arrogância nossa tentar julgar a conduta de outras pessoas. Deveríamos adotar uma atitude de tolerância em relação às práticas de outras culturas.

Rachels, todavia, aponta que tais argumentos são falaciosos. Segundo o autor, o erro básico dessa linha de raciocínio é tentar deduzir uma conclusão sobre o assunto a partir de que diferentes culturas diferem drasticamente em suas perspectivas de certo e errado. Entretanto, ao examinar as diferenças culturais, descobrimos que elas não se diferenciam tanto entre si.

Um exemplo é a proibição do ato de matar, pois encontramos essa regra vigente em todas as sociedades viáveis. Conforme ele leciona:

> As culturas podem se diferenciar no que elas consideram como as exceções legítimas às regras, mas essa divergência ocorre sob um pano de fundo de convenções em questões mais amplas. Portanto, é um erro superestimar o número de diferenças entre as culturas. Nem toda regra moral pode variar de uma sociedade para outra.[6]

Portanto, ele conclui que existem regras morais que a todas as sociedades devem ter em comum, pelo fato de serem inerentes à sua própria existência. As diferenças gerais não estão nos valores, e sim nas crenças e costumes, que podem ser atribuídas a outros aspectos sociais, mas que não alteram drasticamente o padrão moral da sociedade.

2.2 Correntes Teóricas Derivativas da Moral e suas Consequências na tomada de Decisão de Veículos

Terminada a classificação de moralidade, passamos agora a projetar o seu conceito na tomada de decisão dos AVs. Como já ilustrado, a produção em massa desses automóveis trará,

[6] RACHELS. *Op cit,* p. 26.

como inevitável consequência, a ocorrência de um acidente em algum momento. Apesar do próprio intuito do veículo ser a diminuição de colisões, tal fato ainda não é garantido, pois até mesmo os veículos projetados pela Uber e Tesla já foram protagonistas de falhas que ocasionaram fatalidades.

Portanto, independente do progresso tecnológico, alguns acidentes ainda serão inevitáveis, envolvendo risco de vidas humanas, seja no carro que dirige sozinho ou fora do mesmo (GOODALL, 2014). No contexto deste acidente, o veículo terá que decidir que ação tomar, e isso implica em definir as potenciais consequências. Quanto à essa decisão, esse trabalho se aprofundará melhor em sua resposta no próximo capítulo, no qual analisará o "Dilema do Bonde", também conhecido como *Trolley Problem*, que consiste em um antigo debate no qual independente da decisão tomada, uma consequência negativa irá ocorrer, cabendo ao leitor refletir acerca da melhor decisão, do seu ponto de vista moral.

A pergunta que permanece para esse subcapítulo, todavia, é: presumindo que é possível inserir regras éticas nos códigos da IA, de quem devem ser essas regras éticas? Afinal, ninguém ainda foi capaz de propor um conjunto geral de princípios para seres humanos reais que seja aceito universalmente sem algum tipo de contestação. Diante dessa indagação, estão explicadas abaixo as principais correntes morais consideradas pelos pesquisadores da área, que servem de alicerce para a tomada de decisão, sendo elas: Utilitarismo, Deontologismo, Relativismo, Absolutismo e Pluralismo.

2.2.1 Utilitarismo

O Utilitarismo é uma corrente normativa que busca a ação mais benéfica, considerado os aspectos positivos e negativos das consequências relativas à decisão tomada. É, em tese, uma espécie de consequencialismo, no qual o tomador da decisão pondera o que de bom e de ruim acontecerá em seguida, para determinar qual ação irá gerar o resultado mais positivo (Ess, 2014). Aplica-se, a partir daí, o Cálculo Felicífico, derivado do Hedonismo, que serve como uma forma de medir a quantidade de prazer em contraste a dor (bem e mal) para guiar a tomada de decisão. No utilitarismo, portanto, os fins justificam os meios.

Diversos exemplos estão presentes na história, no qual durante uma guerra, o sacrifício de alguns soldados é justificado para proteger uma nação. Nos tempos modernos, espera-se que

os robôs tomem ações utilitaristas, que sacrificariam uma pessoa em prol do bem comum, ou até mesmo sacrificando sua própria existência (MALLE *et al*, 2015).

Do ponto de vista moral, o utilitarismo considera como a melhor ação aquela que produz o melhor resultado. Portanto, no cenário hipotético de uma situação fatal em que a escolha reside entre sacrificar cinco pedestres ou os dois passageiros de um veículo, a decisão do carro autônomo provavelmente seria (assumindo que o número de vidas perdidas é o único critério) matar os passageiros do carro, ao invés dos pedestres.

Dessa forma, em se tratando de um AV projetado com uma IA utilitarista, seria esperado que a decisão tomada buscaria minimizar a quantidades de vidas perdidas, em prol da maximização da felicidade comum.

2.2.2 Deontologismo

O Deontologismo, por outro lado, é uma teoria aplicada por Kant e criada em 1834 por Jeremy Bentham, que deriva do grego *deon* (dever, obrigação) + *logos* (ciência). Ela parte do princípio de que existem regras absolutas que não podem ser quebradas. Os deontologistas, dessa forma, entendem que não são as consequências das ações que importam, mas sim a essência da ação em si e de que forma ela será tomada. Ao seguir essa corrente, portanto, é esperado que se faça a "coisa certa pela razão certa porque é a coisa certa a se fazer" (KARNOUSKOS, 2020).

Os exemplos mais concretos disso estão presentes em algumas religiões, nas quais a vida é sagrada e absoluta, sendo qualquer forma de violência contra a mesma descartada. Com isso, em uma tomada de decisão fatal em um AV, um religioso rejeitaria a ideia de matar um outro humano, não importando as consequências disso, seja até para salvar um grupo de pessoas ou a si mesmo (ESS, 2014). A justificativa para isso se pautaria no argumento de que essas pessoas "não conseguiriam viver sabendo que são responsáveis de alguma forma pela morte de outras pessoas" (FRISON *et al*, 2016).

Conforme mencionado, o deontologismo é aplicado por Kant, através do imperativo categórico "aja apenas com base na máxima através da qual você pode, ao mesmo tempo, desejar que se torne uma lei universal" (KANT, 2001). Isso acaba por determinar quais são os

deveres morais que devemos ter como base, tornado os mesmos regras são absolutas que devem ser obedecidas independente das consequências. Nessa corrente, a ênfase é colocada no próprio ato, em contraste com o utilitarismo, em que a ênfase está no resultado da ação.

Devido a sua natureza absoluta, a deontologia pode ser alvo de críticas, uma vez que prioriza sempre o certo sobre o bem. De acordo com Kant, por exemplo, mentir é sempre errado, mesmo que algo bom possa ser obtido a partir disso. No entanto, essa posição deontológica, especialmente quando o ato de salvar vidas está envolvido, se mostra falho.

Portanto, em se tratando de um AV projetado com uma IA deontologista, a visão que predominaria seria a de tratar todas as pessoas como iguais (evitando atribuir uma valoração à individualidade da pessoa, como sugere o utilitarismo), tomando a melhor decisão com "boas intenções" independente do resultado no caso de um acidente.

2.2.3 Relativismo

Seguindo para a próxima corrente, temos o Relativismo, que se vale do argumento de que todas as normas e valores são validos somente em relação à uma cultura ou um grupo de pessoas. Karnoukos explica que o Relativismo ético se pauta no argumento de que, embora exista uma miríade de práticas consideradas eticamente aceitáveis em algumas sociedades, essas mesmas ações são condenadas em outras. Portanto, aqui é proposto que as práticas de tal sociedade somente podem ser julgadas pelo seu próprio código moral.

Ess (2014) entende que podemos extrair duas vantagens dessa corrente de pensamento, sendo elas: (i) tolerar as visões e práticas de um grupo distinto, uma vez que, para eles, tais ações existem e são aceitáveis e; (ii) oferecer um conforto, uma desculpa, no sentindo de que se tudo é relativo para uma cultura ou grupo específico, então não é necessário buscar exaustivamente por valores globalmente válidos. Quanto ao segundo item, tal argumento já foi debatido no subcapítulo anterior, e provado como falácia, vez que valores morais universais já foram constatados e, portanto, por mais que exista um certo grau de relatividade desses valores em diferentes culturas, o ser humano possui um padrão ético.

O principal problema do Relativismo seria, segundo Karnoukous, a atribuição da moralidade estar atrelada à cultura de um grupo, pois isso acabaria ocasionando uma espécie

de tolerância à determinadas ações pois tal prática foi feita observando o código de ética exclusivo desse grupo. Isso acabaria levando à uma paralisia moral, na qual nenhum julgamento ético poderia ser feito contra a moralidade de um grupo distinto. Por exemplo, não poderíamos julgar o apartheid na África do Sul, ou até mesmo o Nazismo, pois esses eram os valores daqueles grupos, o que demonstra uma grave falha de se encarar o Relativismo como teoria moral absoluta.

Um AV projetado com uma IA Relativista se comportaria de maneira diferente de acordo com os códigos morais da sociedade em que ele está sendo conduzido. Como não existe uma moral universal de certo ou errado para essa teoria, ele se basearia simplesmente no que o meio em que ele está inserido tolera, mesmo que tal ato possa ser condenado em outros lugares do mundo.

2.2.4 Absolutismo

O Absolutismo é o exato oposto do que o Relativismo propõe. Segundo essa corrente de pensamento, existem valores universais de regra, normas, crenças e práticas que definem o que é certo e bom para todos, em todos os momentos e em todos os lugares e os que diferem estão errados. Com isso, tudo o que for de acordo com a visão do absolutista está correto, enquanto o contrário é condenável.

Isso acaba por gerar um confronto interessante de idealizações, na medida que dois absolutistas podem discordar, o que acaba gerando duas posições na qual um vê a outra como errada. O Relativismo, por exemplo, não iria condenar nenhuma das posições, mas considerar e tolerar as diferenças entre ambas, entendendo que essa discordância se dá com base na diferença ética gerada por uma cultura ou grupo social.

Um AV projetado com uma IA Absolutista iria se submeter às regras globais mundiais, e tomaria sua decisão com base nessas regras, independentemente de seu proprietário e condutor concordar com tal ação. Como o direito à vida é um dos maiores pilares do Direito, quase como se fosse algo sagrado, o veículo não poderia decidir em nenhuma hipótese qual vida deveria ceifar, mesmo que seja em prol do bem comum.

2.2.5 Pluralismo

Por fim, o Pluralismo é uma corrente moral que rejeita o Absolutismo e o Relativismo, propondo, como o próprio nome diz, uma pluralidade de verdades morais. Segundo Karnoukous, os pluralistas defendem que existem valores morais universais (Absolutismo), entretanto, entendem que há uma miríade de interpretações possíveis em diversos contextos, podendo ser entendido de diferentes maneiras (Relativismo). As diferenças aqui podem ser toleradas, mas existem valores morais básicos, que se infringidos, são caracterizados como imorais.

Ess preceitua que o pluralismo ético nos permite ver como as pessoas em diversas culturas podem compartilhar normas e valores importantes, mas, ao mesmo tempo, são capazes de interpretar e aplicar essas normas e valores em outros espectros, que refletem diferentes contextos históricos e culturais.

Talvez essa corrente moral seja a mais difícil de implementar em uma IA, visto que a pluralidade na moral implica que cada veículo teria uma maneira de tomar decisões diferentes, a depender do seu proprietário. Em um cenário Pluralista, a Tesla, por exemplo, fabricaria dois modelos de um AV, sendo um dotado de padrões mais utilitarista e um mais egoísta. Em uma situação dilemática, o veículo utilitarista sacrifica seu dono para causar o menor prejuízo, enquanto o egoísta faz tudo o que pode para salvar seu dono, mesmo que isso signifique matar os dois garotos, por exemplo (HARARI, 2018).

3. AS BARREIRAS ÉTICAS NA TOMADA DE DECISÃO DE UM VEÍCULO AUTÔNOMO

3.1 O Dilema do Bonde (The *Trolley* Problem)

As teorias morais explicadas no capítulo anterior incidem no mais infame "problema sem solução" já pensado pela humanidade, conhecido como o Dilema do Bonde, internacionalmente chamado de *The Trolley Problem*. No contexto dos AVs, frear abruptamente para evitar bater em uma pessoa ou animal pode o passageiro do veículo em perigo, e os fabricantes desses automóveis terão que implementar essas escolhas em seus algoritmos, podendo às vezes tomar decisões de vida ou morte em cenários mais extremos, o que se assemelha com o Dilema do Bonde.

Elaborado por Philippa Foot, o dilema envolve um bonde desgovernado que está prestes a cruzar um trilho que contém cinco pessoas amarradas, que, eventualmente, serão mortas. Nós, no papel de observador, temos a possibilidade de puxar uma alavanca, que desviaria o curso do bonde para a direita, salvando os cinco indivíduos ali amarrados. Entretanto, nos trilhos da direita existe um único homem, que agora será morto, em detrimento dos outros cinco.

O principal objetivo de Foot, quando criou esse dilema, era apresentar um debate acerca da doutrina de duplo efeito, utilizada por católicos para discutir acerca do aborto. Nas palavras de Renato Kinouchi:

> 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 órgãos para outros cinco pacientes[7].

Para David Edmonds (2014), a doutrina do duplo efeito pode ter uma fórmula objetiva, contendo quatro componentes:

- O ato que visamos executar, se retirado as consequências que serão causadas, não pode ser moralmente errado (no dilema do bonde, por exemplo, o ato visado é salvar uma vida, e a consequência é a morte).

- O agente pretende o bem e não visa causar o dano tanto como meio ou fim, embora consiga prever que o dano será causado.

- Não há como alcançar o resultado positivo (bem) sem causar efetivamente o dano.

- Os efeitos negativos não são desproporcionalmente grandes em relação ao bem que está sendo buscado.

Edmonds, portanto, faz uma analogia de que podemos justificar um ataque à uma base terrorista utilizando esses quatro preceitos. Afinal (1) atacar uma base terrorista não é um ato moralmente errado. (2) O ataque por si só é o ato pretendido, não o dano colateral que irá ser causado. (3) É impossível atacar a base terrorista sem causar os danos colaterais. (4) Os efeitos colaterais do ataque não podem ser desproporcionais ao ataque à base (como, por exemplo, destruir uma cidade inteira).

No *Trolley Problem*, qual seria, portanto, a decisão correta: puxar a alavanca ou deixar o bonde seguir seu caminho? Tanto puxar a alavanca quanto permanecer inerte são ações que se enquadram na doutrina de duplo efeito, na medida em que o objetivo final do ato é salvar vidas, mas haverá uma consequência que não pode ser evitada, mas não é de pretensão do agente.

---

[7] KINOUCHI, Renato. Sobre as Limitações do Dilema do Bonde para a Avaliação dos Riscos Impostos por Veículos Autônomos. *Revista de Filosofia Moderna e Contemporânea*, v. 6, pp. 115-130, 2018.

A grande maioria das pessoas acredita que não só é permitido puxar a alavanca, como é quase necessário – moralmente obrigatório. Entretanto, essa decisão não é uma resposta objetiva que pode ser usada como regra moral, afinal, o dilema aqui presente é objeto de um intenso debate filosófico que perdura há décadas e possui diferentes visões de teorias morais.

Ambas as respostas parecem ser plausíveis e justificáveis. Em seus estudos, Foot chegou à conclusão de que o dever negativo de se abster de executar a ação "ruim" de puxar a alavanca prevaleceria sobre o dever positivo de puxar e salvar uma vida.

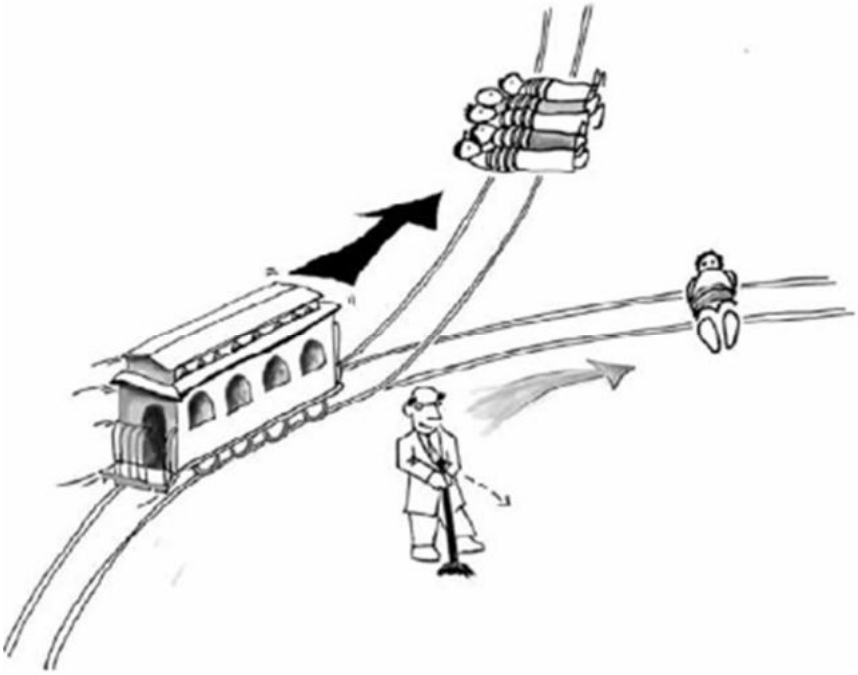

Figura 2 – O Dilema do Bonde. Temos a opção de puxar uma alavanca para salvar a vida de cinco pessoas. Entretanto, tal ato implica na morte do homem à direita. Qual opção escolher? Fonte:Google

Um utilitarista, por exemplo, chegaria facilmente à conclusão de que a alavanca deve ser puxada, pois considera todos os aspectos positivos e negativos da decisão tomada. Como o aspecto positivo de salvar cinco vidas é maior que o aspecto negativo de ceifar uma, a decisão fica clara, pois produz o melhor resultado. O argumento do utilitarismo aqui, serie que uma ação é moralmente aceita se o resultado visar a maximização da "felicidade" para o maior

número de indivíduos (AWAD, 2017). Porém é moralmente correto escolher a vida de cinco pessoas em troca de uma?

Por sua vez, um deontologista que segue a teoria Kantiana encara a eticidade de cada ação sob uma ótica universalista, em que existem regras absolutas que não podem ser quebradas e que o respeito à dignidade, como critério absoluto, demonstra ser errado usar os seres humanos em prol do bem-estar geral, como os utilitaristas pregam (SANDEL, 2016). Podemos fundamentar que mudar o trilho de direção constitui efetivamente um ato de matar uma pessoa, enquanto não fazer nada significa meramente permitir que alguém morra (nesse caso, cinco mortes), sendo moralmente pior matar do que deixar morrer (LIN, 2015). Portanto, a aplicação da teoria de Kant ao dilema do bonde resultaria em uma inação, o que ocasionaria a morte de cinco pessoas pois, a partir do momento em que se opta por salvar as pessoas, estaríamos quebrando uma regra absoluta por meio de uma tomada de decisão.

Matar implica que somos diretamente responsáveis pela morte de uma pessoa: se não tivéssemos feito tal ato a pessoa estaria viva. Deixar morrer, no entanto, envolve muito menos responsabilidade de nossa parte, uma vez que o processo causal já estava em andamento e não foi iniciado ou controlado por nós.

Todavia, soa justa a morte de cinco pessoas em contrapartida à sobrevivência de uma? Percebe-se que a ética kantiana também não serve como verdade absoluta. Saber se é pior matar do que deixar morrer é debatido no campo da filosofia há anos, e engloba diversos argumentos sem, entretanto, se chegar a um consenso universal e objetivo.

Para desenvolver ainda mais o debate acerca desse problema moral, Judith Thomson (1976, 1985) propôs acrescentar uma nova variável ao dilema. Nesse novo cenário, nos encontramos em uma passarela observando os trilhos e vemos, novamente, o bonde vindo para colidir com cinco pessoas amarradas aos trilhos.

Existe, porém, um homem gordo debruçado sobre a passarela observando o bonde vindo e, se empurrarmos esse homem ele cairá em cima dos trilhos. Thompson informa que o homem é tão obeso que o seu tamanho faria o bonde parar e salvar as cincos pessoas, em troca, entretanto, de sua vida.

Diante disso, seria aceitável se incumbir do dever negativo de empurrar o homem almejando o dever positivo de salvar as cinco vidas? O principal ponto deste debate acerca do *Trolley Problem* se dá na medida em que na sua primeira variante, a maioria das pessoas julga como admissível puxar a alavanca que desvia o curso do bonde, ocasionando a morte da pessoa no trilho da direita; em contraste apenas uma minoria julga aceitável impulsionar o homem gordo para sua morte (GREENE, 2013).

O objetivo de Thomson (2008) nessa proposta era justamente demonstrar esse tipo de assimetria em nosso julgamento moral. Isto é, por que é permitido em um caso salvar as cinco vidas sacrificando uma, enquanto no outro caso não?

Frances Kamm aprofunda essa indagação, preceituando que o problema filosófico básico advindo dos dilemas é: por que certas pessoas, usando certos métodos, têm permissão moral para matar menos pessoas para salvar um número maior, enquanto outras, usando diferentes métodos, não têm permissão moral para matar o mesmo número menor para salvar o mesmo número maior (KAHM, 2015)?

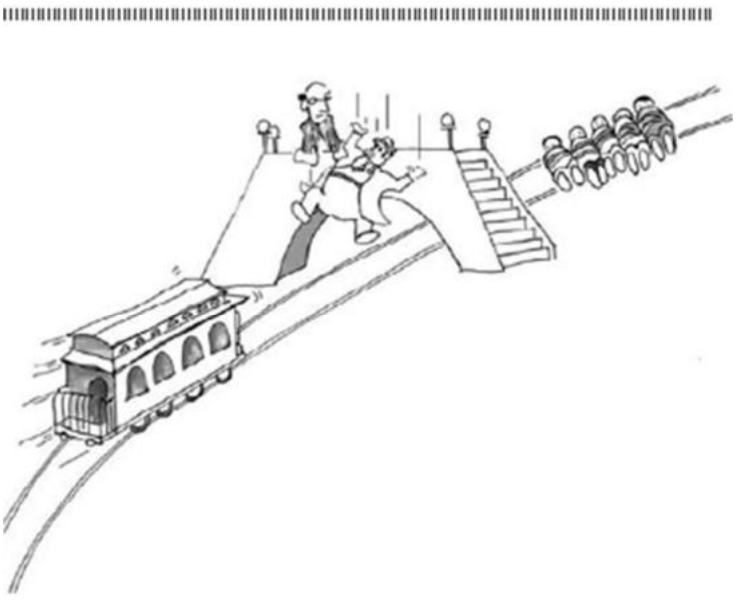

Figura 3 – O dilema com o homem gordo. Temos a opção de empurrar o homem para parar o curso do bonde. Por que a grande maioria das pessoas evita tomar essa ação? Fonte: Google

Joshua Greene (2013), um renomado neurocientista de Harvard, afirma que o Dilema do Bonde desencadeia uma furiosa luta cerebral entre as faces calculistas e emocionais do cérebro. Confrontados pelo problema do homem gordo e a opção de matarmos com nossas

próprias mãos, partes do cérebro associadas a sentimentos como compaixão (amígdala cerebelosa e córtex pré-frontal) entram em estado de atividade frenética, desencadeando um "alarme emocional" em nosso cérebro que nos faz pensar que tal atitude é errada por questões inteiramente morais. Sem esse "alarme emocional", entretanto, as atividades cerebrais são registradas nas partes mais racionais (córtex pré-frontal dorsolateral e lobo parietal inferior), o que nos faz tender a uma ação que utilize um cálculo utilitarista, equacionando na balança diversos prós e contras que não esbarram unicamente no campo da moralidade (EDMONDS, 2014).

Greene afirma que o gatilho emocional que normalmente ocorre quando as pessoas contemplam matar o homem gordo é composto de dois componentes, sendo eles i) a fisicalidade do ato de empurrar (isto é, utilizar nossos músculos para impactar diretamente outra pessoa) e; ii) a intencionalidade de causar prejuízo à uma pessoa.

Visando elucidar esses dois pontos, Greene (2016) conduziu um estudo, adaptando o Dilema do Bonde para distintas situações, visando entender a interação entre utilização de força física e intenção no julgamento da moralidade. Em um de seus experimentos, ele constatou que mesmo sem contato físico direto com as mãos (o ato de empurrar em si), se utilizarmos nosso corpo para empurrar o homem gordo de alguma maneira, como, por exemplo, utilizando uma vara longa, a aceitação moral ainda é baixa, já que, nesse caso, temos presentes a intencionalidade e o ato de empurrar.

É o caso do *Trap Door Dilemma*, exemplo no qual nós temos a opção de salvar as cinco vidas ao puxar uma alavanca que irá abrir um buraco abaixo do homem gordo, derrubando-o nos trilhos e parando o bonde. Apesar do resultado envolver a morte do homem gordo, menos pessoas estariam dispostas a empurrar diretamente o homem se comparado a abrir a armadilha, comprovando, dessa forma, o aspecto da fisicalidade. Ainda assim, seja empurrando ou ativando a armadilha, a grande maioria das pessoas ainda acredita que matar o homem gordo é pior do que mudar a direção do bonde (EDMONDS, 2014).

Entramos, dessa maneira, no segundo fator proposto por Greene, no qual ficamos relutantes em causar prejuízo a um terceiro intencionalmente, em concordância com aquilo que preceitua a doutrina do duplo efeito, tratada anteriormente neste capítulo. Como nesse caso temos a intenção de causar a morte do homem gordo como um meio para o resultado de salvar

cinco pessoas, estamos menos propensos a cometer tal ato do que mudar a trajetória do bonde e matar alguém como meramente efeito colateral.

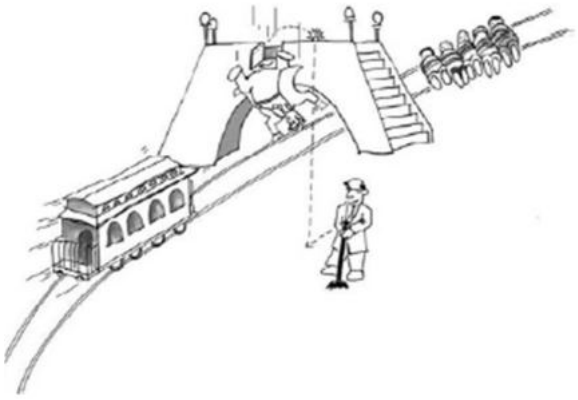

Figura 4 – O dilema da armadilha. A única maneira de parar o bonde é puxar uma alavanca que derrubaria o homem gordo e o mataria, mas salvaria as cinco pessoas. Qual a diferença entre empurrar e puxar a alavanca? Fonte: Google

Greene conduziu, com isso, diversas variâncias do dilema do bonde e concluiu, com os resultados obtidos através de respostas de participantes em seus experimentos, que ações que causam um dano prejudicial são moralmente menos aceitas conforme o nível de fisicalidade. Além disso, a fisicalidade interage com a intenção de modo que o fator de força pessoal afeta apenas o julgamento moral de danos pretendidos, *i.e*, puxar a alavanca fisicamente para desviar o percurso do trem não causa um julgamento moral negativo (GREENE, 2016). Ele conclui, dessa forma, sugerindo que o nosso senso de moral de uma ação está atrelado diretamente as propriedades motoras mais básicas e que o fator da intenção está intimamente ligado à nossa sensibilidade de fisicalidade.

Podemos estabelecer, portanto, que apesar de um grupo maior de pessoas se sujeitar a tomar a decisão com menos fisicalidade e intenção no dilema aqui debatido, existe ainda uma parcela que escolheria jogar o homem gordo da ponte para salvar cinco pessoas, o que evidencia a falta de consenso quanto à decisão moralmente correta e o complexo dilema moral que os fabricantes dos AVs devem se ater.

3.2 Críticas ao Dilema do Bonde aplicadas à Tomada de Decisão de Veículos Autônomos

Apesar do Dilema do Bonde ser um assunto formidável para debater moralidade e investigar a diferença entre deveres positivos e negativos e o ato de matar e "deixar morrer", alguns autores entendem que existem diferenças cruciais em se fazer uma direta analogia entre a tomada de decisão de veículos e o dilema em si. Dado que os experimentos éticos que o *Trolley Problem* propõe são análogos aos dilemas imaginados para carros autônomos, é razoável assumir que os resultados morais que a sociedade escolhe para solucionar o Dilema do Bonde são suficientes para resolver as mesmas situações éticas se enfrentadas na vivência da realidade?

Nassim Jafarinaimi defende, nesse caso, que o Dilema do Bonde é insuficiente para resolver a discussão ética acerca da tomada de decisão de AVs. Conforme ela leciona:

> Considere, por exemplo, o fato de que não aprendemos quaisquer detalhes sobre os cinco trabalhadores no enquadramento original do dilema do bonde: Quem são eles? Eles são jovens ou velhos? Como eles estão posicionados em relação ao bonde? Eles são capazes de ver, ouvir e reagir ao bonde conforme sua aproximação? Estaria eu, o motorista do bonde, mirando injustamente em um trabalhador, visto que ele é apenas um espectador da situação até que eu decida direcioná-lo, dando-lhe muito menos tempo para reagir? Essas e outras questões semelhantes apontam para a natureza incerta, complexa e viva de a situação que eu, o motorista do bonde, poderia estar enfrentando. Como um resultado, embora possa parecer que o princípio utilitarista de "salvar a maioria das vidas" é aplicável quando apresentado com a versão simplificada do cenário, este princípio pode ou não servir em uma situação real (tradução nossa)[8].

É certo que a IA dos AVs será testada na posição de tomar uma decisão em frações de segundos que virão a ter implicações na vida ou morte de pessoas (WALLACH, 2009). Entretanto, essa decisão já estará definida em um estágio anterior, na fase de treinamento do

---

[8] JAFARINAIMI, N. *Our Bodies in the Trolley's Path, or Why Self-driving Cars Must Not Be Programmed to Kill*. Science, Technology, & Human Values. 2018. Texto original: "*Consider, for example, the fact that we do not learn any details about the five workers in the original framing of the trolley problem: Who are they? Are they young or old? How are they positioned in relationship to the trolley? Are they capable of seeing, hearing, and reacting to the trolley as it approaches them? Would I, the trolley driver, be unfairly targeting the one worker, given that he is just a bystander in the situation until I decide to target him, thus giving him a lot less time to reflect and react? These and other similar questions point to the uncertain, complex, and living nature of the situation that I, the driver of the trolley, could plausibly be facing. As a result, while it may seem that the utilitarian principle of "saving most lives" is applicable when presented with the simplified version of the scenario, this principle may or may not serve in an actual lived situation."*

algoritmo do sistema do veículo, quando será decidido de que forma os AVs serão "programados" para responder aos cenários de acidentes. Acerca desta etapa de programação, entretanto, existe um problema conhecido como *Black Box Problem*, que será aprofundado no próximo capítulo.

Em total contraste, no Dilema do Bonde estamos em face de um iminente acidente, sem tempo nenhum para pensar. Devemos decidir ali mesmo o que fazer: puxar a alavanca para desviar o bonde ou empurrar o homem gordo? À título de exemplo, um motorista de um carro convencional pode repentinamente enfrentar uma situação em que precisa decidir, naquele exato momento, se deve atropelar uma pessoa para salvar cinco. Essa sim é uma situação análoga ao *Trolley Problem*, em contraste ao dilema que os AVs deverão enfrentar (NYHOLM e SMIDS, 2016).

Sven Nyholm & Jilles Smids (2016) argumentam, ainda, que a decisão tomada pelo AV é fruto de diversas "cabeças pensantes", em contraste com a decisão de uma única pessoa no Dilema do Bonde. Nesse caso, a tomada de decisão da IA é realisticamente representada como sendo feita por cidadãos comuns, advogados, engenheiros, especialistas em ética, fabricantes de automóveis etc., que irão chegar em uma solução mútua, de comum acordo.

Sven Hansson (2012) introduz, ainda, uma crítica relacionada à exclusão da questão de exposição a riscos nas discussões éticas acerca do Dilema do Bonde, na medida em que na situação hipotética, sabemos com total certeza que cinco pessoas irão morrer caso permaneçamos inertes, contrariamente à morte de uma, se agirmos.

Isso é um claro contraste a um dilema real, visto que os problemas humanos não são acompanhados de um conhecimento certo sobre suas consequências. Por exemplo, se aplicarmos o *Trolley Problem* de maneira crua à uma tomada de decisão de IA, podemos imaginar um caminhão que está em rota de colisão com um AV tripulado por cinco pessoas, sendo a única maneira de salvar os tripulantes uma guinada que irá ceifar a vida de um pedestre (NYHOLM e SMIDS, 2016). O Dilema do Bonde implica que independente da decisão tomada, alguém irá morrer com total certeza. Entretanto, não há como garantir que alguém virá a óbito ou sobreviverá com ferimentos leves, o que torna o problema um pouco mais complexo, com a adição de alguns elementos (GODDAL, 2014).

Nyholm & Smids exemplificam isso ao questionar detalhes como o peso do caminhão, a possibilidade de que o motorista freie ou tente desviar o trajeto e até mesmo se os cinco passageiros estão usando cinto de segurança. Todos esses fatores contribuem para as chances de sobrevivência, em total contraste ao dilema original.

Dessa forma, podemos obter um modelo do dilema do bonde probabilístico, no qual podemos considerar o enunciado: um bonde sem freios está em iminência de atropelar cinco pessoas com probabilidade de 10% para que elas morram. Um observador se encontra próximo de uma alavanca, tendo a possibilidade de acioná-la e desviar o bonde para um outro trilho, atropelando outra pessoa cuja probabilidade de morte seja de 50% (KINOUCHI, 2018). Com as probabilidades definidas, podemos calcular situações de risco em que a IA do veículo melhor poderá decidir, baseado nas melhores chances matemáticas.

Essa linha de argumentação parece ser um ponta pé inicial eficaz para a solução moral do dilema debatido de um ponto de vista utilitarista, em que o menor prejuízo seria causado e não arriscaríamos uma colisão baseada no número de pessoas atingidas e sim usando diversos fatores para calcular um risco probabilístico e chegar ao melhor resultado.

A probabilidade de risco, inclusive, tem sido o caminho adotado pelos desenvolvedores de AVs, tendo como desafios a mapeação de condições a serem programadas e determinar o grau de confiabilidade aceitável no cálculo de risco para a realização de ações (GOODALL, 2016).

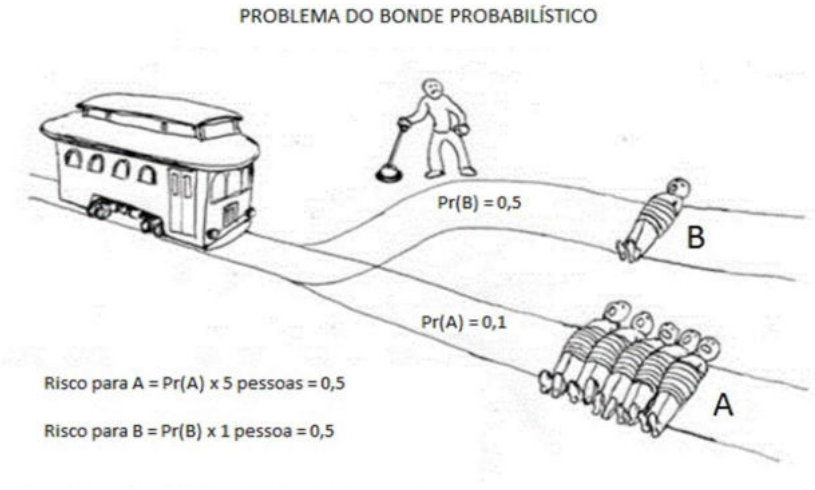

Figura 5 – O Dilema do Bonde probabilístico, no qual podemos chegar matematicamente por meio da gestão de risco a uma solução que cause os menores prejuízos. Fonte: Kinouchi (2018)

Fica evidente, dessa forma, que o dilema do bonde não deve ser considerado como o modelo perfeito para se resolver as questões éticas relacionadas a tomada de decisão de AVs por não traduzir a realidade da situação enfrentada, na medida em que o dilema se mostra simplório em demasia, sem considerar os fatores aqui apresentados. Portanto, este subcapítulo buscou demonstrar três críticas ao *Trolley Problem* que evidenciam as diferenças entre este dilema e a tomada de decisão da IA de um AV, podendo essas críticas serem resumidas em:

- Quanto à tomada de decisão, no Dilema do Bonde, temos apenas uma pessoa responsável, enquanto na IA de AV essa decisão foi previamente debatida por um grupo de indivíduos.

- Quanto ao tempo necessário para a tomada de decisão, no Dilema do Bonde a ação é imediata, sendo necessária uma ação quase que instantânea, ao contrário da IA do veículo, que já tem um plano de contingência para o enfrentamento do evento.

- Quanto ao conhecimento da situação, o Dilema do Bonde implica que independente da ação tomada, o resultado será conhecido e exato, já na tomada de decisão da IA, haverá incerteza quanto ao resultado da ação, devendo ser feita uma estimativa de risco para minimizar os prejuízos.

3.3 *Moral Machine*: As Escolhas Éticas e a Ausência de Consenso Social sobre o Sacrifício

Conforme debatido no subcapítulo anterior, o dilema do bonde, por considerar apenas duas soluções possíveis com consequências determinadas, não corresponde a realidade em se tratando de uma situação a ser vivida por um AV. Entretanto, transpor a situação do dilema para um AV e realizar uma pesquisa para averiguar a percepção da sociedade acerca dos dilemas morais no caso de colisões inevitáveis pode servir de parâmetro na discussão entre fabricantes, juristas e governo (BONNEFON; SHARIFF; RAHWAN, 2015).

Para determinar o consenso social diante da decisão de um AV, Jean-François Bonnefon, Azim Sharrif e Iyad Rahwan conduziram seis estudos online, com 1928 participantes no total, contendo situações dilemáticas. No primeiro estudo, 182 participantes foram entrevistados e 76% deles foram favoráveis a ideia de que um AV sacrifique um passageiro ao

invés de matar dez pedestres. Perguntados quanto maneira “mais moral” de se programar um AV em uma escala de 0 (proteger o passageiro a todo custo) a 100 (minimizar o número de casualidades) os pesquisados demonstraram tendências a priorizar veículos utilitaristas, minimizando o prejuízo quantitativo de vidas, com uma média de 85 encontrada (BONNEFON; SHARIFF; RAHWAN, 2016).

Em prosseguimento a esta ideia, no segundo estudo, com 451 pessoas, os participantes foram confrontados com um número de pedestres a serem salvos, variando de 1 a 100, em detrimento do passageiro do AV, e, logicamente, quanto mais vidas eram salvar, maior a aprovação social, chegando aos mencionados 76% do primeiro estudo quando 10 vidas deveriam ser salvas.

No terceiro estudo, que contou com 259 participantes, os membros da pesquisa foram apresentados a uma situação em que um membro de suas famílias estaria no AV, e foram perguntados, em uma escala de 1 a 100, o quão inclinados estariam a comprar um AV utilitarista, isto é, que sacrificaria suas vidas e de seu familiar. Em total contraste ao primeiro estudo, que resultou num consenso de que AVs utilitaristas eram a opção mais moral, a média encontrada foi de 19, o que demonstrou que, com um familiar na equação, as pessoas tendem a preferir veículos que protejam os passageiros.

O quarto estudo, com 267 participantes, pediu que os entrevistados alocassem 100 pontos para indicar (i) o quão moral os algoritmos do AV que seria apresentado eram; (ii) o quão confortável os participantes estavam para que outros AVs fossem programados da mesma maneira e; (iii) o quão propensos estavam para comprar um AV programado com os algoritmos do item (i). O AV apresentado estaria em três situações distintas, sendo elas: (i) atropelar um pedestre para salvar dez pessoas; (ii) matar o próprio passageiro para salvar dez pessoas e; (iii) matar um pedestre para salvar outro pedestre.

O primeiro algoritmo logicamente recebeu o maior número de pontos, em concordância com o que primeiro e segundo estudo concluíram. Não obstante, o segundo algoritmo recebeu pontos mistos no quesito moralidade (cerca de 50), entretanto, no quesito de intenção de compra, recebeu apenas 33 pontos de média.

Em resumo, esses estudos indicam que, apesar dos participantes serem a favor de uma IA utilitarista em seus AV (em que seria moralmente correto o sacrifício de seus passageiros em prol de menos casualidades) e terem disposição para comprarem esses veículos, quando questionados com a possibilidade de estarem pessoalmente envolvidos no dilema (ou seus familiares), a propensão de compra drasticamente diminui, fazendo com que os participantes demonstram interesse em veículos que protegem os passageiros. Dessa forma, parece que os passageiros preferem continuar com um condutor humano a um AV que possa vir a priorizar outras vidas em detrimento da sua (FLEETWOOD, 2017).

Este estudo, entretanto, não analisou as probabilidades de resultados na tomada de decisões, considerando, conforme o *Trolley Problem*, que a situação resultaria em uma fatalidade, não havendo uma decisão que gerasse um menor risco, como por exemplo um AV decidindo se chocar contra um muro, havendo a chance de sobrevivência de seus passageiros.

Em 2018, os pesquisadores voltaram a fazer estudos com cidadãos, argumentando que as decisões sobre os princípios éticos que nortearão os AVs, não podem ser deixadas apenas para os engenheiros ou especialistas em ética. Para que os consumidores mudem dos carros tradicionais para os AVs e para o público em geral aceitar sua presença nas ruas, ambos precisarão compreender as origens dos princípios éticos que são programados nestes veículos (BONEFFON et al, 2017). Em outras palavras, mesmo que os especialistas em ética chegassem a um consenso sobre como AVs devem resolver seus dilemas morais, o resultado seria inútil se os cidadãos discordassem da solução encontrada, evidenciando, dessa forma, que qualquer tentativa de conceber um padrão ético na inteligência artificial deve estar em consonância com a moralidade da sociedade (AWAD, 2018).

Isso evidencia, portanto, a necessidade de um entendimento refinado de como diferentes indivíduos e países podem diferir em suas preferências éticas se quisermos progredir em direção à utópica ética universal para IA de um AV (GRAHAM et al, 2016). Dessa forma, foi designado o *Moral Machine*, no qual usuários se depariariam com acidentes inevitáveis com dois resultados possíveis, dependendo se o AV desvia ou permanece no curso. A diferença entre os estudos citados ao longo desse capítulo para o *Moral Machine* reside no fato de que neste último, diversas variáveis foram inseridas, como, por exemplo, preservar animais domésticos, crianças, idosos e até pessoas com condições econômicas precárias.

No total, 492,921 pessoas participaram do experimento, englobando 233 países e, em linhas gerais, o pesquisadores constataram uma preferência quase que universal em preservar a vida de (i) humanos em detrimento de animais domésticos, (ii) um maior grupo de pessoas (evidenciando novamente a tendência utilitarista humana frente à dilemas morais) em relação a um grupo menor de pessoas e; (iii) os mais jovens em detrimento dos mais velhos.

Isso iria, entretanto, de encontro com as regras éticas propostas em 2017 pelo *German Ethics Comission on Automated and Connected Driving* (Comissão Alemã de Ética sobre Direção Automatizada e Conectada), mais especificamente a regra de número 9, que estabelece que qualquer distinção com base em características pessoais, como por exemplo idade, deve ser proibido.

O mais interessante, todavia, reside no fato de que usando a geolocalização, os pesquisadores puderam mapear em que locais do mundo as preferências eram mais fortes, agrupando os países em grupos que possuíam escolhas morais similares.

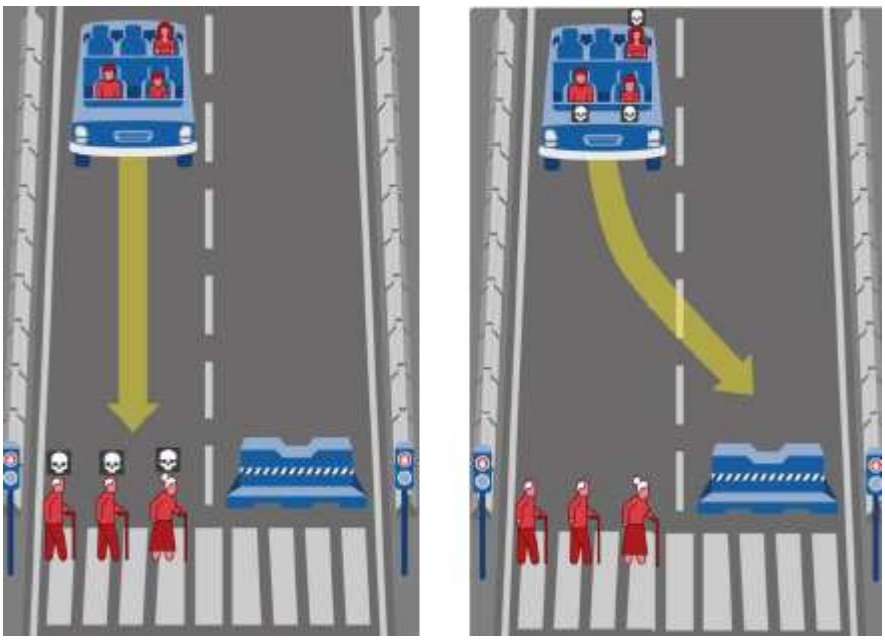

Figura 6 – A interface do *Moral Machine*. Um AV com defeito em seu freio está em iminente colisão. Se o veículo permanece em linha reta, irá ceifar a vida três idosos, sendo dois homens e uma mulher. Caso vire a direita, irá bater num muro e ceifar a vida de seus passageiros: um casal de adultos de sexo oposto e um menino.
Fonte: AWAD (2018)

Por exemplo, a preferência por poupar os jovens em relação aos mais velhos se mostra muito mais presente no grupo do Sul (que consiste nos países da América Central e do Sul, além de França e países com influência francesa) do que no grupo do Leste (que inclui países asiáticos e nações do Oriente Médio). Já a preferência por preservar humanos sobre animais domésticos era mais fraca no grupo do Sul se comparada com o grupo Leste e Oeste (este que é constituído dos Estados Unidos, Canadá e grande parte da Europa).

Chegou-se à conclusão, dessa forma, que as preferências resultantes do experimento do *Moral Machine* estão correlacionadas às variações culturais e econômicas dos países. No caso dos países do grupo Oriental, como o respeito aos mais velhos é algo que está inserido de forma mais forte em sua cultura, a tendência de preferir poupar os mais jovens ser mais fraca consegue ser facilmente explicada.

Outra importante questão para os reguladores considerarem é saber se os pedestres estavam seguindo a lei (ex: atravessar corretamente no sinal fechado). Devem aqueles que estão atravessando a rua de maneira incorreta gozar da mesma proteção que o pedestre que atravessa corretamente (AWAD 2018)? O estudo concluiu que a prosperidade econômica do país influencia diretamente na aceitação desse quesito. Traduzindo, participantes de países mais pobres e com instituições menos presentes são mais tolerantes com os pedestres que cruzam a rua sem obedecer às leis, possivelmente por causa de suas experiências com menor cumprimentos de regras e punições mais fracas em seu sistema.

Para Thomas Pölzler (2021), a previsão é que os carros autônomos se estabelecerão de forma mais rápida em nossa sociedade conforme eles estão em conformidade com nossas crenças morais. Por exemplo, uma pessoa que acredita que os VAs devem atropelar um adulto ao invés de uma criança estará mais inclinado a comprar, aprovar a compra e apoiar a legislação em favor de VAs se esse carro efetivamente tiver algoritmos que tendem a proteger crianças e atropelar adultos. Este efeito provavelmente será ainda mais forte no caso de decisões morais que envolvem a própria segurança das pessoas como motoristas pois as pessoas, é claro, preferem ser salvas (BONNEFON et al., 2016).

Ou seja, Pölzler defende a teoria do Relativismo cultural, exemplificada no capítulo 3 desta monografia. Caso os VAs adotassem o Absolutismo (Universalismo), também descrito no capítulo anterior, eles poderiam divergir do que as pessoas considerassem eticamente certo,

o que faria a sociedade ter uma menor tendência em querer adquiri-los. Ele argumenta, ainda, que como os VAs serão fabricados em poucos países e exportados, uma IA Universalista iria instaurar uma espécie de colonialismo moral, no qual o VA regido pelo algoritmo de uma pequena proporção da humanidade impõe seus pontos de vista morais sobre o resto do mundo.

Portanto, a programação de AVs com uma IA Universalista corre o risco de gerar resultados negativos nas sociedades e culturas que terão padrões éticos impostos a elas. Isso pode causar indignação pública, causar hostilidade em relação aos países ou fabricantes que impõe os algoritmos do AV e, o mais importante, pode impedir que algumas sociedades e culturas introduzam legislações visando a redução de carros autônomos, perdendo ou atrasando o aumentos de bem-estar que esses carros tendem a gerar. Ao permitir que cada sociedade ou cultura determine seu próprio algoritmo de acidente, o AV Relativista evita esses efeitos negativos.

Entretanto, quanto a ideia de o AV Relativista ser a escolha moral mais correta, Pölzler afirma que eles podem ter consequências moralmente ultrajantes. Podemos supor que exista uma crença dominante em uma sociedade que, em casos de acidentes, os AVs devem sempre favorecer os caucasianos sobre as pessoas de cor, ou os homens sobre as mulheres. Sob o pressuposto do Relativismo, isso significaria que os carros autônomos devem ser moralmente programados para refletir essas crenças. Mas algoritmos de acidente operando com base nessas preferências morais não seriam bem-vistos pelo mundo, apresentando um contraponto a ideia do AV Relativista.

Em uma entrevista recente, Udo Di Fabio, um ex-juiz constitucional e líder de uma comissão alemã sobre a ética de carros autônomos, expressou preocupação sobre a possibilidade de a China treinar AVs para poupar pessoas com pontuação alta em seu sistema de crédito social. Fato é que as pessoas detentoras de maior poderio econômico poderiam subornar autoridades ou fabricantes de AVs de modo a aumentar a sua própria segurança ou de suas famílias nas rodovias (POLZLER, 2021) Existe, portanto, argumentos plausíveis que denotam os contras de um AV Relativista e, para mitigar isso, leis de direitos humanas ou estaduais/federais deverão prever a proibição deste tipo de discriminação.

Fica evidente, dessa forma, que existem preferências culturais quanto às questões éticas, conforme explanado no capítulo 3 desta monografia, no que concerne a Teoria Moral do

Relativismo. O fato de que a ampla maioria das regiões do mundo demonstrou uma alta taxa de concordância com relação às decisões éticas escolhidas sugerem que a IA dos AVs deve estar alinhada com esses valores, do contrário, os reguladores irão ter desafios para explicar a razão pela qual os AVs não seguem decisões morais consideradas como corretas pela sociedade, além de enfrentar uma repercussão negativa.

4. EVIDENCIANDO AS DIFICULDADES NA REGULAÇÃO DE RESPONSABILIDADE NA TRANSIÇÃO PARA A AUTOMAÇÃO DE VEÍCULOS

Passada a evidenciação das preferências morais da sociedade, permanece ainda um desafio quanto a evidenciação da integração dos AVs em nosso cotidiano. Integrar uma IA em um veículo é uma tarefa que exige cautela, conforme demonstrado com os dilemas apresentados ao longo dessa monografia. Entretanto, como não podemos responsabilizar os robôs ou uma IA por suas ações, os legisladores e os tribunais precisarão determinar quais partes serão responsáveis por acidentes, além de determinar como compensar adequadamente aqueles que sofreram algum prejuízo. Se a solução legislativa proposta não for justa, os fabricantes de IA podem acabar recebendo menos investimento, pois os riscos podem não compensar.

Neste capítulo serão tratados os sujeitos potencialmente responsáveis pelos acidentes automobilísticos envolvendo IA, abordando de que forma as suas responsabilidades deverão mudar com provável transição do veículo tradicional para um VA. Os sujeitos aqui analisados serão os fabricantes, os motoristas e, por fim, os pedestres.

4.1 Responsabilidade dos Fabricantes no Desenvolvimento de um Veículo Autônomo Seguro

Como o fabricante do AV deve ser legalmente responsável de alguma forma quando um defeito acontecer no veículo ou um erro vier a ser encontrado em seu programa, a sociedade precisará decidir como impor a responsabilidade pelo acidente. No cenário atual, a imensa maioria dos acidentes de carro é resultante de um erro de motorista e não envolve um defeito no veículo (ZOHN, 2015). Entretanto, com a automação, é esperado que acidentes sejam causados.

A responsabilidade por acidentes causados por um motorista humano é regida pela responsabilidade subjetiva, e no caso de culpa, nas modalidades da negligência, imprudência e imperícia (SHROLL, 2015). No entanto, quando um defeito em um veículo é a causa responsável pelo acidente, a questão jurídica se complica, na medida em que essas modalidades de culpa somem e transformam o caso em uma responsabilidade do fabricante pelo produto (ANDERSON *et al*. 2016).

Jeffrey Gurney (2017), aponta uma possível implicação séria, pois poderia transformar um setor já estabelecido de jurisprudência (acidentes de carro) em uma área completamente nova, no qual seria necessário avaliar a responsabilidade da fabricante do veículo sempre que houver com acidente. Gurney nos apresenta dois exemplos, visando ilustrar de que maneira os processos judiciais poderiam ser dificultados. Em suas palavras:

> Considere dois acidentes diferentes. No primeiro acidente, Ken está dirigindo um veículo tradicional. Ele adormece, fazendo com que seu veículo desvie para uma faixa de tráfego contrária e colida com o veículo de Elaine. Aqui, a responsabilidade pode ser resolvida de maneira simples: Ken seria o responsável; sua seguradora provavelmente pagaria até os limites de sua apólice; e um processo seria evitado. Mesmo se a seguradora não pagasse, o caso poderia ser decidido facilmente no tribunal. Ken tinha o dever de prestar atenção à estrada e dirigir dentro de sua própria pista; ele violou esses deveres quando adormeceu e seu veículo desviou para a faixa de tráfego em sentido contrário; sua violação causou danos a Elaine; e Elaine sofreu prejuízos. No segundo acidente, Ken possui um veículo autônomo que foi produzido pela Montadora ABC. Ken comprou o veículo depois de ver um anúncio pela ABC, que afirmava que seus veículos autônomos permitiam que os ocupantes se envolvam em atividades além de observar a estrada. Durante seu trajeto para casa após um longo dia de trabalho, Ken tira uma soneca; e, enquanto ele está dormindo, seu veículo autônomo entra em uma pista de sentido contrário e atinge o veículo de Elaine. Apenas ferimentos leves e ocorrem pequenos danos à propriedade (tradução nossa)[9].

Gurney explica que, de acordo com as circunstâncias apresentadas na situação, Ken não foi negligente. Portanto, se não foi, Elaine deveria resguardar seus direitos contra a fabricante ABC, o que, judicialmente falando, é muito mais difícil do que obter sua compensação de Ken.

---

[9] GURNEY, Jeffrey K. Imputing Driverhood, Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles" *In* LIN, Patrick *et al*. *Robot Ethics 2.0:* From Autonomous Cars to Artificial Intelligence. Oxford: Oxford University Press, 2017. Texto original: "*Consider two different accidents. In the first accident, Ken is driving a traditional vehicle. He falls asleep, causing his vehicle to drift into a lane of oncoming traffic and strike Elaine's vehicle. Here, liability could be resolved simply: Ken would be responsible; his liability insurer would likely pay up to his policy limits; and a lawsuit would be averted. Even if the insurer did not pay, the case could be decided easily in court. Ken had duties to pay attention to the road and to drive within his own lane; he breached those duties when he fell asleep and his vehicle drifted into the oncoming traffic lane; his breach caused Elaine's harm; and Elaine suffered damages. In the second accident, Ken owns an autonomous vehicle that was produced by ABC Manufacturing. Ken purchased the vehicle after seeing an advertisement by ABC that claimed its autonomous vehicles allow occupants to engage in activities other than watching the road. During his commute home after a long day at work, Ken takes a nap; while he is sleeping, his autonomous vehicle drifts into the lane of oncoming traffic and strikes Elaine's vehicle. Only minor injuries and minor property damage occur."*

Se tratado como um caso de defeito de produto pela justiça, Elaine não poderia usar o fato de que o carro lhe causou um dano para provar que o fabricante do veículo teria responsabilidade, devendo provar que houve algum defeito ou negligência na produção do veículo (GURNEY, 2017). Provar este tipo de defeito, por si só, demanda tempo e dinheiro, fazendo com que ingressar com esse tipo de ação não valha a pena no final, pois os custos poderiam se igualar à compensação monetária. Evidente então a complexidade que um processo de acidentes envolvendo um AV pode vir a ter em relação a um veículo comum.

Para tentar contornar esse eventual problema, Gurney sugere que os legisladores poderiam tratar o fabricante do veículo como o próprio motorista do carro. Se o fabricante for considerado o motorista, no caso do exemplo dado anteriormente, a Montadora ABC tinha o dever de manter seu veículo na pista correta e o quebrou quando invadiu a faixa de Elaine, simplificando a parte contenciosa. A indagação que fica, entretanto, é se poderemos equiparar fabricantes de AVs à motoristas de veículos. Segundo Gurney, essa alternativa é provavelmente o que a sociedade espera que os legisladores apliquem, estando em linha com as expectativas das partes lesadas e dos fabricantes em geral.

Para verificar a responsabilidade nesses casos de equiparação, Ujjayini Bose (2015) sugere que o AV tenha sempre uma caixa preta, semelhante às aeronaves, que registre os eventos. Essas caixas pretas fornecerão tribunais - ou, nesse caso, companhias de seguros - as informações necessárias para determinar a causa do acidente. Usando essas informações, o tribunal ou o júri poderia determinar se o fabricante "dirigiu" o AV de maneira razoável (BOSE, 2015). Entretanto, como veremos mais a frente, a caixa preta pode não ser o suficiente para averiguar a responsabilidade do fabricante.

Todavia, nos casos em que o proprietário do veículo queira ajuizar uma ação, transportando essa situação para o Brasil, entendemos que não aplicaríamos o Código Civil (que exigiria comprovação do defeito no VA pela parte lesada), mas sim o Código de Defesa do Consumidor (CDC), utilizando a teoria do risco do empreendimento e atividade empresarial, preceituando que todos que se disporem a exercer qualquer atividade no mercado de consumo tem obrigação de responder por eventuais vícios ou defeitos do produto. Portanto, a responsabilidade pelo fato do produto, conforme explicada no art. 12 do CDC, terá como fato gerador o defeito do produto.

Como a posição do CDC é favorável à vítima no caso de acidentes decorrentes de relações consumeristas, por considerá-la a parte mais frágil da relação de consumo e hipossuficiente, o fabricante do VA terá que provar que seu produto não possui defeito, vez que domina o processo produtivo e, para a vítima, é oneroso em demasia produzir alguma prova técnica ou científica do defeito em juízo.

Passado esse tópico evidenciando possíveis embaraços ao acesso à justiça nos casos de acidente com AV, iremos separar a responsabilidade dos fabricantes em três grupos, a ver:

4.1.1 Responsabilidade mecatrônica (*Duty of Care*)

Para o primeiro tipo de responsabilidade tratada, iremos analisar os casos em que um defeito ocorre no veículo, mas especificamente nos componentes elencados no capítulo 2 deste trabalho, na fase de sensoriamento. Quando um motorista causa um acidente devido a um erro, ele próprio é o culpado pelo dano (ANDERSON et al, 2016). Este fato, entretanto, muda quando um veículo se pilota sozinho, já que não será mais a pessoa sentada atrás do volante a responsável pela direção.

Os AVs são projetados para serem controlados por um complexo sistema que utiliza radares, lasers, lidares, camêras de vídeo e posicionamento por GPS (Duffy & Hopkins, 2013). Aqui, parece ser razoável assumir que os fabricantes continuarão a assumir responsabilidade pelo desempenho das tecnologias de sensores, motores e outros componentes. Conforme entende Lim (2018), deve haver um *duty of care* (dever de cuidado), que, de forma ampla, se traduz na obrigatoriedade que os fabricantes possuem de que seus produtos sejam seguros ao exercer a função para a qual foram criados.

Dessa forma, o *duty of care* que recai sobre o fabricante está ligado intrinsicamente ao pleno funcionamento individual dos componentes listados no segundo capítulo desta monografia (apesar de que esse dever não se limita apenas a esses componentes), e a forma que eles funcionam em conjunto para detectar todos os empecilhos no trajeto e evitar, dessa forma, qualquer tipo de colisão. Com todos os sensores funcionando harmonicamente e sem nenhum defeito, pode se argumentar que o fabricante atingiu satisfatoriamente o dever de cuidado.

Há que se mencionar, ainda que o *duty of care* também abarca possibilidade de que o AV consiga dar ao condutor alguma forma de controle do veículo, em situações emergenciais.

Um veículo no Nível 5[10], por exemplo, deveria conter um volante e pedais de freio e aceleração, caso um imprevisto aconteça. Uma vez que ocorra uma quebra do *duty of care*, o condutor teria como, manualmente, tentar controlar a situação.

À título de exemplo, podemos destacar o acidente do Tesla Model S em 2016, a primeira colisão envolvendo um AV que resultou em uma fatalidade humana. O acidente consistiu em um veículo de nível 2 sendo pilotado por sua IA, que acabou colidindo com um caminhão enquanto virava para a esquerda em um cruzamento em razão de uma falha no sistema de frenagem. A quebra do *duty of care* aqui foi verificada quando se constatou, posteriormente, que as câmeras desse modelo não possuíam tecnologia que detectavam de forma eficaz quando um veículo se aproximava perpendicularmente, o que fez com que o sistema de freios não fosse acionado da maneira que se espera.

4.1.2 Responsabilidade pelo software da Inteligência Artificial

A responsabilidade pelo sistema de inteligência artificial nada mais é que o debate suscitado por essa monografia no que tange à decisão dos veículos autônomos frente à uma situação de acidente. Atualmente, quanto aos acidentes com veículos comuns, os tribunais se baseiam em uma análise das decisões que as pessoas envolvidas no acidente tomaram, bem como as condições carro no momento do acidente. No caso dos AVs, entretanto, essas decisões terão que ser determinadas baseadas no algoritmo do programa presentes nos veículos em momento pretérito ao acidente.

Isso significa que na aferição de responsabilidade, os tribunais deverão considerar a "previsão" dos fabricantes quanto aquele evento, em vez de uma decisão tomada em tempo real. Isso implica na necessidade de um sistema que averigue a precisão e confiabilidade dos dados em tempo real no carro, tal como uma caixa preta, conforme explicitada no início desse capítulo. No processo de desenvolvimento de um software, a imprevisibilidade de todos os possíveis defeitos é algo concreto, e, um grande problema é que, o reparo desses possíveis erros induz à modificação das linhas de código, o que ocasionar novos erros, em um efeito "bola de neve".

[10] Conforme já explicado no início desse trabalho, o veículo Level 5 seria aquele que o condutor não teria nenhuma espécie de controle, pois ele teria suas funções integralmente automatizadas.

Nesse interim, a teoria da computação defende que inexiste um programa de computador em que se garanta isenção de possíveis defeitos. Afinal, é impossível exigir do ser humano que avalie as infinitas possibilidades geradas pelas linhas de código do algoritmo (LEVY, 2001). Para se resguardar de tais falhas, seria necessário utilizar um programa que verifique os erros, o que, por si só, já demonstra um paradoxo, visto que o próprio programa de verificação pode conter erros, exigindo que o mesmo seja examinado (e quem vai verificar o verificador?).

Tais defeitos se relacionam com o que os especialistas definem como "risco de desenvolvimento", e inexiste consenso mundial quanto ao assunto. Atualmente no Brasil podemos notar duas correntes distintas, sendo elas: (i) favorável à ideia do risco de desenvolvimento como excludente de responsabilidade do fabricante e; (ii) a favor da responsabilização do fornecedor. Seria, entretanto, um defeito imprevisível uma atitude ilícita que violaria uma norma jurídica?

Em nosso ordenamento jurídico atual, a IA e os robôs não podem ser responsabilizados por ações ou omissões que causam danos a terceiros, justamente por serem destituídos de personalidade jurídica (ALBIANI, 2018). As normas existentes indicam que a falha do robô pode ser atribuída a um agente humano, como o fabricante ou até mesmo o utilizador (em casos que o comportamento poderia ser de alguma forma mitigado). Isso demonstra que as normas atuais não estão prontas para receber a tecnologia das IA, visto que os danos provocados serão feitos por uma capacidade adaptativa de aprendizagem, via *machine learning*, que possuem um comportamento caracterizado como imprevisível, vez que aprendem de acordo com sua experiência e muitas vezes de maneira autônoma (HALLEVY, 2010).

Na Europa, inclusive, artigos já estão sendo publicados que concluem pela impossibilidade de se atribuir uma responsabilidade civil à uma personalidade jurídica da IA dos robôs, visto que tal personalidade é fictícia e não se assemelha ao caráter humanista pela qual a União Europeia se pauta, podendo levantar futuramente o risco de diminuir o *status* do ser humano ao se nivelar com uma IA (NEVEJANS, 2016).

A falha em processar os dados conforme o algoritmo manda é um problema que assola as IAs dos AVs, sendo uma situação análoga a falha de um componente, e com o registro na caixa, poderíamos entender o que levou àquela tomada de decisão.

A caixa preta que registraria os dados referentes à tomada de decisão dos AVs, entretanto, não é tão fácil de ser desvendada. Esse é um problema que os cientistas chamam de *Black Box Problem* e, para ser explicado, faz-se necessário, antes, uma breve explanação do conceito de *machine learning*.

Machine learning nada mais é do que uma aplicação da inteligência artificial que fornece ao sistema a habilidade de automaticamente aprender e melhorar com base em experiências sem precisar ser explicitamente programada, ou seja, podem acessar dados e aprender por si. O processo de aprendizagem começa com observações ou dados, como instruções, experiências ou até mesmo exemplos, a fim de encontrar padrões nos dados e tomar melhores decisões no futuro com base nos exemplos que fornecemos. O objetivo principal é permitir que os computadores aprendam automaticamente sem intervenção humana ou assistência e ajustar as ações de acordo (EXPERT AI, 2020).

Para que a IA esteja apta a responder problemas realísticos e não só um número limitado de cenários, o *Machine Learning* usa uma quantidade exorbitante de dados para aprimorar seus métodos de decisão e planejamento, correspondendo ao fenômeno do *big data* mencionado no segundo capítulo deste trabalho.

Dentro do *Machine Learning*, existem algumas técnicas utilizadas pelos desenvolvedores de IA para aperfeiçoar o aprendizado do sistema. Podemos destacar, resumidamente:

- *Reinforcement Learning*: é um método que interage com o ambiente, produzindo ações e descobrindo erros ou recompensas. Este método permite que softwares determinem o comportamento ideal dentro de um contexto específico para maximizar seu desempenho, sendo necessário um feedback de recompensa para que a IA entenda qual a melhor ação, ato conhecido como sinal de reforço.

- *Supervised Learning*: aplica o que foi aprendido no passado a novos dados, usando exemplos denominados para “prever” eventos futuros. O programa é treinado sobre um conjunto de dados pré-definidos. Baseado nesse treinamento, o programa pode tomar decisões precisas quando receber novos dados.

- *Deep Learning*: os algoritmos do *deep learning* podem ser considerados uma evolução sofitsticada, pois analisam dados com uma lógica similar a como um humano infere conclusões. Para conseguir isso, a IA usa uma estrutura em camadas de algoritmos chamada de *Artificial Neural Network* (ANN). O design da ANN é similar a uma rede neural de um cérebro humano, resultando em um processo de aprendizagem muito mais eficiente do que nos modelos tradicionais de *machine learning*.

Voltando ao *Black Box Problem*, seria possível "abrir" a caixa preta do AV e descobrir por que a IA tomou tal decisão? Essa indagação, entretanto, carece de resposta. Isso porque o grande problema do *Black Box* é justamente não conseguirmos explicar o que as matrizes de algoritmos representam. Entendemos que a caixa preta funciona medindo a precisão dos dados de teste, mas não sabemos definir como.

O óbice, portanto, reside no fato de que os sistemas de IA possuem uma opacidade inerente: sabemos que os *inputs* vão resultar em um *output* algorítmico, em uma decisão. No entanto, como o algoritmo chegou ao resultado específico – quais partes dos dados de *input* o algoritmo considera importante para prever um resultado – muitas vezes não é transparente (GUIDOTTI *et al*, 2019).

Hoje, com a gigantesca rede neural de algoritmos implementada nas IAs e com o *deep learning* treinando e usando vastos arquivos de *big data*, é quase impossível desvendar esse mistério. Em uma eventual ação judicial, provar que uma decisão foi tomada por meio de um erro na IA do AV é impraticável. Mesmo se o ônus da prova, com uma eventual responsabilidade objetiva, recaísse sobre o fabricante do veículo, explicar a decisão tomada pelo AV no meio de uma miríade de linhas de código é impossível (LIM, 2018).

Existe também uma possível preocupação que, através do *deep learning*, os algoritmos se tornem discriminatórios, afinal, eles são construções humanas: criados, treinados e programados por pessoas. Isso implica que as falhas derivadas do raciocínio humano e sua tomada de decisão podem ter certa influência sobre a aplicação de algoritmos.

Muitas das correlações que os algoritmos detectam podem ser considerados resultado de discriminação histórica e preconceitos no nosso raciocínio. É sabido, por exemplo, que historicamente os homens eram nomeados para cargos de alta função e, se um algoritmo fosse

usado para filtrar e detectar candidatos para serem contratados baseado na chance de eles serem promovidos, por conta desse fato histórico, o algoritmo provavelmente selecionaria apenas candidatos do sexo masculino. Dessa forma, mesmo que o algoritmo pareça estar simplesmente sendo "honesto", para evitar discriminação, será exigido que rejeitemos esses resultados (GERARDS, 2019).

Por isso, se um algoritmo não for cautelosamente programado ou se for treinado com dados tendenciosos, há um grande risco de que visões subjetivas sejam alimentados a ele. Por exemplo, em uma IA que auxilia um juiz no julgamento de um caso, o uso de algoritmos para se chegar em uma sentença mais adequada pode carregar visões humanas tendenciosas e condenar pessoas baseadas em sua condição financeira, visto a tendência de impunidade para aqueles que possuem maior capital financeiro. Por causa do *black box problem*, o juiz não conseguiria detectar essa inclinação algorítmica e, mesmo ciente do problema, o algoritmo forneceria a ele jurisprudências ou doutrinas que inclinariam o pensamento humano a concordar com a decisão obtida.

Uma importante tentativa de solucionar esse dilema é fazer com que os algoritmos que governem a IA sejam "*accountable*". *Accountability* é ser obrigado a explicar e justificar uma ação ou omissão, devendo responder a alguém (OLSEN, 2014). Bovens (2007) define *accountability* como uma relação entre uma pessoa e um júri, na qual a pessoa tem a obrigação de explicar e justificar sua conduta e o júri pode fazer perguntas e julgar para que, dependendo da resposta, a pessoa enfrente consequências.

Nesse contexto, esses déficits informacionais decorrentes da inerente opacidade e complexidade dos algoritmos tem implicações diretas no que tange a responsabilidade prevista na *accountability*, isto é, de serem explicáveis ou justificáveis (BUSUIOC, 2021). Explicar o porquê da IA tomar determinada decisão é justamente a dificuldade, e, sem poder entender a razão, indivíduos afetados adversamente por decisões algorítmicas não podem contestar as decisões tomadas.

Para então poder traçar uma linha de responsabilidade quanto às decisões de uma IA, os legisladores devem, primeiramente, entender como os algoritmos funcionam. Entretanto, o problema reside, primeiramente, no fato de que esses algoritmos são protegidos como segredos comerciais, visto que o *source code* de um programa pode ser equiparado à "formula secreta" de uma empresa, e seu vazamento implicaria em um possível fortalecimento de concorrentes e,

eventualmente, em uma perda de lucros. O artigo 2º da *EU Trade Secrets Directive* (Diretiva de Segredos Comerciais da União Européia) estabelece que, para garantir proteção, a informação (no caso, o algoritmo) deve i) ser secreta; ii) ter valor comercial devido ao seu sigilo; e iii) estar sujeita a medidas para manter o seu segredo.

Entretanto, essa proteção não é absoluta, visto que o Recital 11 da *EU General Data Protection Regulation* (Regulamento Geral de Proteção de Dados da União Européia) estabelece que tal sigilo deverá ser ignorado caso as regras de um país ou da União exigirem que os detentores de segredos comerciais divulguem, por razões de interesse público, informações – incluindo tais segredos – ao público ou às autoridades administrativas ou judiciais para desempenho de suas funções. Todavia, inexistem diretrizes para ajudar os juízes ou autoridades públicas a traçar o ponto de equilíbrio em casos de conflito entre o interesse privado de manter um algoritmo em segredo e o interesse público (HUSEINZADE, 2021).

Contudo, mesmo se imaginarmos casos em que o sigilo acima descrito seja derrubado por autoridades, e um órgão regulador consiga exigir que o fabricante de uma IA deva deixar o código de programação disponível para eventuais fiscalizações e testes, isso traria um fardo excessivo para as autoridades, pois se incumbir de testar milhões de linhas de código parece ser inviável.

Para Lim (2018) é inconcebível fiscalizar com precisão a área de IA, pois demandaria uma equipe altamente especializada e fundos excessivos, o que torna tal prática inviável para os cofres públicos. Além disso, seria quase que impossível exigir que o órgão regulador medisse a qualidade do algoritmo via *machine learning*, devido à necessidade de se alimentar o veículo com uma quantidade massiva de *big data*, para que ele, efetivamente, acabe "aprendendo".

A solução talvez seja que reguladores exijam que a comunidade científica desenvolva modelos de IA e *machine learning* compreensíveis e interpretáveis em um primeiro momento, e que a indústria adote sistematicamente práticas que simplifiquem o entendimento dos softwares.

Este subcapítulo, portanto, buscou evidenciar a inerente dificuldade em se tratar questões jurídicas advindas da tomada de decisão envolvendo uma IA.

4.1.3 Responsabilidade pela proteção de dados

Dado que a coleta, processamento, armazenamento e transferência de dados desempenham funções vitais dos AVs, a segurança de dados também deverá ser discutida como uma possibilidade de responsabilização em caso de quebra. Uma grande preocupação que pode surgir, por exemplo, advém da possibilidade de se aferir a localização de uma pessoa através do dispositivo de GPS. Em 2012, a Suprema Corte dos Estados Unidos entendeu que instalar um dispositivo GPS em um veículo e rastrear sua localização constitui uma quebra do *Fourth Amendment*, e não pode ser conduzida sem um mandado (US v. JONES, 2012).

Ao saber a localização de uma pessoa, podemos descobrir diversas informações pessoais, o que demonstra a necessidade de medidas de segurança rigorosas para que um AV não as passe sem necessidade. Alguns AVs, inclusive, permitem a integração de redes sociais para que o carro forneça maiores funções automatizadas, como por exemplo ligar para um amigo ou escutar determinada música. Apesar de ser conveniente, essas informações nas mãos erradas podem ser perigosas e, caso vazadas, devem ser passíveis de responsabilização por parte dos fabricantes.

A pior hipótese, entretanto, envolve cyber segurança e cyber terrorismo. Os dados gerados pelo AV não são apenas vulneráveis a roubos, mas também podem ser alvo de sequestro por parte de sujeitos mal-intencionados. Se o veículo estiver conectado à uma rede para baixar informações, por exemplo, existe um ponto de acesso que pode ser violado e usado para cortar informações críticas ou até mesmo assumir o controle do veículo. É o caso de um modelo Tesla X, por exemplo, que foi hackeado e conseguiram controlar a abertura de portas, teto solar, ligar e desligar luzes e até frear[11].

Cabe aos órgãos fiscalizadores pressionar os fabricantes para que esses sistemas sejam virtualmente impenetráveis. Propomos aqui que a imputação de responsabilidade nesses casos deve ser mais severa, visto os perigos advindos desses ataques, devendo recair na modalidade objetiva.

---

[11] GREENBERG, Andy. This Bluetooth Attack Can Steal a Tesla Model X in Minutes. 23 nov. 2020. Disponível em: https://www.wired.com/story/tesla-model-x-hack-bluetooth/. Acesso em: 25 maio 2021.

Isso denota, portanto, que há uma crescente necessidade de proteção de direitos fundamentais contra entidades privadas que manuseiam algoritmos, visto que os mecanismos legais da nossa sociedade não estão preparados para lidar com as violações de direitos fundamentais em relações horizontais (GERARDS, 2019).

4.2 Responsabilidade dos Motoristas na Transição para os Veículos Autônomos

Embora o objetivo dos AVs seja eliminar completamente o papel do motorista, há muitas etapas no período de transição até a sua real implementação – estamos atualmente no nível 3 – que vão exigir que o condutor humano ainda seja necessário. Com o crescente nível de automação, riscos gerados pela direção distraída (como por exemplo utilizar o celular ou mandar mensagens enquanto o veículo está em movimento) devem aumentar exponencialmente. À medida em que mais funções do veículo se tornem automatizadas, a tendência é que motoristas se tornem dependentes delas, negligenciando potenciais deficiências tecnológicas.

Diante disso, é de interesse dos fabricantes dos AVs que o legislador responsabilize os motoristas pela eventual falta de atenção, mesmo em situações que o veículo dirigindo sem interferência humana. Entretanto, evidente que o motorista humano é uma fraca medida de segurança para quando o sistema autônomo do veículo falhar.

À título de exemplo, podemos evidenciar o caso de Joshua Brown, que estava em seu Tesla *Model S* no piloto automático quando um caminhão passou em sua frente. A forte luz solar refletida no branco do caminhão confundiu a tecnologia de visão computacional do Tesla, o que fez o veículo não ser rápido o suficiente para frear e evitar um acidente. As funções automatizadas do veículo poderiam ser, a qualquer momento, sobrescritas, mas Brown não conseguiu reagir para desviar ou pressionar o freio antes da colisão (KURTZ, 2016).

Importante salientar que, fazer com que um condutor humano concorde com uma lei que vise imputar responsabilidade a ele, não deverá isentar os fabricantes das responsabilidades advindas de um design seguro de veículo. Entendemos que, mesmo nos casos em que um defeito de projeto não é a causa direta de um acidente, se tal defeito colocar em uma pessoa e aumentar os danos durante um acidente, o fabricante ainda deverá ser responsável. O desafio jurídico aqui será averiguar os limites das responsabilidades do motorista e do fabricante nos

casos em que esses acidentes tenham origem em algum defeito mas o condutor deveria e poderia ter agido.

Dessa forma, tendo em vista o atual estágio de desenvolvimentos dos AVs, é necessário que os fabricantes comuniquem claramente as limitações de suas tecnologias, além de investir em funções que mitigam o ato da "direção distraída". Da mesma maneira, os motoristas deverão estar cientes dessas limitações, utilizando os mecanismos projetados para assumir controle do veículo quando necessário.

Concluímos que os litígios pós-acidente irão desempenhar um importante papel na atribuição de responsabilidade entre o motorista e o fabricante, devido à complexa natureza de se determinar se a necessidade de constante atenção e iminente resposta podem ser esperadas de um humano quando o veículo não desempenha corretamente a sua função de piloto automático.

4.3 Responsabilidade dos Pedestres na Transição para os Veículos Autônomos

Os pedestres evidentemente continuarão compartilhando as ruas com os veículos, mesmo na transição proposta da direção autônoma. Entendemos que o principal desafio aqui virá do fato de que os pedestres não seguem as regras de trânsito exatamente como estão escritas, na medida em que aderem, de fato, práticas sociais que já estão enraizadas. O *Moral Machine*, inclusive, já discorreu sobre esse fato, evidenciando que diferentes culturas aderem à jeitos divergentes de transitar nas ruas, e, consequentemente, existe uma diferença no nível de aceitação social quanto a possibilidade de atropelamento de pessoas que não estão seguindo as regras de trânsito.

No Brasil, por exemplo, atravessar uma rua fora da faixa de pedestres não é visto como um comportamento condenável, apesar de ferir o artigo 254 do Código de Trânsito Brasileiro e ser considerada uma infração, com aplicação de multa. É possível que os veículos autônomos enfrentem extremas dificuldades para transitar em locais com desvios de comportamento humano. Em um acidente envolvendo a Uber, um AV atingiu um pedestre que cruzada uma estrada, no Arizona. O veículo, apesar de estar totalmente equipado com tecnologia projetada para evitar colisões e com um motorista pronto para assumir o controle em caso de emergência,

acabou atropelando o pedestre e o matando, que havia atravessado fora da faixa correta (ROBERTS, 2018).

O acidente narrado acima levanta questões sobre a responsabilidade advinda de um AV em casos de travessias abruptas de pedestres. Seria necessário que os pedestres mudem seus comportamentos e a eles seja incumbida a responsabilidade de se adaptar aos AVs? Uma coisa é certa: devido a natureza aleatória do comportamento humano, a IA dos AVs deverá ser projetada com algoritmos que treinem via *deep learning*, em contraste à algoritmos codificados que seguem estritamente a lei, se quisermos compartilhar pedestres e veículos autônomos no mesmo ambiente.

CONCLUSÃO

Conclui-se, portanto, que esse trabalho buscou evidenciar os desafios que a sociedade irá passar com a inevitável chegada dos veículos autônomos no cotidiano social. Com o iminente desenvolvimento da inteligência artificial, o debate ético e jurídico acerca desse tema é essencial, sendo crítico seu amadurecimento na sociedade brasileira, que ainda está em fase inicial.

A criação de padrões éticos na tomada de decisões de uma inteligência artificial, conforme a *German Ethics Comission on Automated and Connected Driving* estabeleceu, é de caráter fundamental para nortear o desenvolvimento da área, devendo ser adotado um código de conduta para o desenvolvimento da IA, para que os algoritmos não desenvolvam comportamentos moralmente divergentes daqueles aceitos pela sociedade, conforme demonstrado no *Moral Machine*, e que operem de maneira ética.

Além disso, destacamos aqui também a imperiosa necessidade de criação de estatutos jurídicos para responsabilizar os sujeitos que farão parte da relação de um veículo autônomo, sendo eles fabricante, condutor e pedestres. Essa regulamentação deve disciplinar com exatidão a conduta dos engenheiros e empresas envolvidas no desenvolvimento da inteligência artificial, sendo eles responsáveis pelos danos causados por sua tecnologia, além de imputar responsabilidades para os condutores que não se mostrarem atentos nos casos em que se faz necessário e pedestres que tenham comportamentos que arrisquem a segurança nas rodovias.

O desenvolvimento da inteligência artificial em veículos autônomos é uma via de mão dupla: não podemos imputar apenas as responsabilidades em cima dos fabricantes, vez que isso iria desencorajar em demasia o avanço tecnológico nesse setor. A sociedade inteira deve fazer sua parte e entender que mudanças serão necessárias. Se tomarmos os cuidados necessários, com certeza será possível desfrutar de todos os benefícios que essa nova tecnologia pretende nos trazer.

REFERÊNCIAS

ABRAMS, Rachel; KURTZ, Annalyn. Joshua Brown, Who Died in Self-Driving Accident, Tested Limits of His Tesla. July, 2016. Disponível em: https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html. Acesso em: 25 maio 2021.

ALBIANI, Christine. *Responsabilidade Civil e Inteligência artificial:* Quem responde pelos danos causados por robôs inteligentes? [s.L]. 2018.

ALLEN, Colin. WALLACH, Wendell. *Moral Machines*: Teaching Robots Right from Wrong. Oxford: Oxford University Press, 2008.

ANDERSON, James M., Nidhi Kalra, Karlyn D. Stanley, Paul Sorensen, Constantine Samaras, and Oluwatobi A. Oluwatola. *Autonomous Vehicle Technology:* A Guide for Policymakers. Santa Monica, CA: RAND Corporation. 2016. Disponível em: http:// www.rand.org/ pubs/ research_ reports/ RR443- 2.html. Acesso em: 25 maio 2021ª

ARAUJO, Valéria Amorim Arantes. Cognição, afetividade e moralidade. *Educação e Pesquisa*, São Paulo, v. 26, n. 2, p. 137-153, 2000.

AWAD, Edmond. *Moral Machine:* Perception of Moral Judgement made by machines. Cambridge: Massachusetts Institute of Technology. [s.L]. 2017. Disponível em: https://damprod.media.mit.edu/x/2017/06/18/awad-ms-17.pdf. Acesso em: 25 maio 2021.

AWAD, Edmond. *The Moral Machine Experiment*. [s.L]. 2018.

B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. Sacrifice one for the good of many: People apply different moral norms to human and robot agents. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human - Robot Interaction. HRI '15. ACM, 2015.

BLOOM, P. *How do morals change?* [s.L]. 2010.

BOVENS, Mark *et al*. *The Oxford Handbook of Public Accountability*. Oxford: Oxford University Press, 2014.

BUSUIOC, Madalina. Accountable Artificial Intelligence: Holding Algorithms to Account. August 2020. Disponível em: https://doi.org/10.1111/puar.13293. Acesso em: 25 maio 2021.

DENG, B. *The robot's dilemma*. Nature, 523. [s.L]. 2015.

EDMONDS, David. *Would you kill the fat man* - the trolley problem and what your answer tells us about right and wrong. [s.L]. 2014.

ESS, C. *Digital Media Ethics*. 2nd ed. [s.L]: Polity Press, 2014.

FLEETWOOD, J. Public Health, Ethics, and Autonomous Vehicles. *American Journal of Public Health*, [s.L], v. 107, n. 4, p. 632-537, 2017.

FOOT, Philippa. *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices*. Oxford: Oxford University Press. 1967.

GERARDS, J. *The fundamental rights challenges of algorithms*. Netherlands Quarterly of Human Rights. [s.L]. 2019.

GERT, Bernard. *The Definition of Morality* (The Stanford Encyclopedia of Philosophy). [s.L], 2020.

GOODALL, N. J. Machine ethics and automated vehicles. *G. Meyer and S. Beiker*, Road vehicle automation, [s.L], Springer, p. 93-102, 2014.

GOODALL, Noah J. Ethical decision making during automated vehicle crashes. *Journal of the Transportation Research Board*, [s.L], vol. 2424, p. 58–65, Dec. 2014.

GRAHAM, J *et al.* Cultural differences in moral judgment and behavior, across and within societies. *Curr. Opin. Psychol.* [s.L], v. 8, 125–130, 2016.

GREENBERG, Andy. This Bluetooth Attack Can Steal a Tesla Model X in Minutes. 23 nov. 2020. Disponível em: https://www.wired.com/story/tesla-model-x-hack-bluetooth/. Acesso em: 25 maio 2021.
GREENE, J. D. Solving the trolley problem. *In* SYTSMA, J. *et al.* A companion to experimental philosophy. [s.L], 2016.

GREENE, Joshua. *Moral tribes:* emotion, reason, and the gap between us and them. New York: Penguin, 2013.

GUIDOTTI, Riccardo *et al*. *A Survey of Methods for Explaining Black Box Models*, 2019.

GURNEY, Jeffrey K. Imputing Driverhood, Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles" *In* LIN, Patrick *et al*. *Robot Ethics 2.0:* From Autonomous Cars to Artificial Intelligence. Oxford: Oxford University Press, 2017.

HALLEVY, Gabriel. *The criminal liability of artificial intelligence entities* – From science fiction to legal social control. Intellectual Property Journal, [s.L]. 2010. Disponível em: http://heinonline.org/HOL/LandingPage?handle=hein.journals/akrintel4&div=11&id=&page. Acesso em: 25 maio 2021.

HARARI, Yuval Noah. *21 lições para o século 21*. [s.L]: Companhia das Letras, 2018.

HUCKO, Filip. *The development of autonomous vehicles.* Copenhagen: Aalborg University Copenhagen, 2017.

HUSEINZADE, Nazrin. Algorithm Transparency: How to Eat the Cake and Have It Too. January 2021. Disponível em: https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/. Acesso em: 25 maio 2021.

JAFARINAIMI, N. *Our Bodies in the Trolley's Path, or Why Self-driving Cars Must Not Be Programmed to Kill*. Science, Technology, & Human Values. 2018.

K. Frison, P. Wintersberger, and A. Riener. First person trolley problem. Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct. Automotive'UI 16. Association for Computing Machinery (ACM), 2016.

KANT, I. *Metafísica dos Costumes*. Paris: Flamarion, 1994.

KANT, Immanuel. Crítica da Razão Pura. 5ª Edição. Lisboa: Fundação Calouste Gulbenkian, 2001.

KARNOUSKOS, S. Ethical frameworks in critical decisions and the acceptance of self-driving cars. Tese de Mestrado. Stockholm University, 2017.

KARNOUSKOS, S. Self-Driving Car Acceptance and the Role of Ethics *In* IEEE Transactions on Engineering Management, [s.L], vol. 67, no. 2, pp. 252-265, 2020.

KINOUCHI, Renato. Sobre as Limitações do Dilema do Bonde para a Avaliação dos Riscos Impostos por Veículos Autônomos. *Revista de Filosofia Moderna e Contemporânea*, v. 6, pp. 115-130, 2018.

LA TAILLE, Yves de. *Moral e ética:* dimensões intelectuais e afetivas. Porto Alegre: Artmed, 2007.

LÉVY, Pierre. *As tecnologias da inteligência*: o futuro do pensamento na era da informática. 10. ed. Rio de Janeiro: Editora 34, 2001.
LIM, Y. H. *Autonomous Vehicles and the Law: Technology, Algorithms and Ethics*. Cheltenham: Edward Elgar Publishing. 2018.

LIN, Patrick *et al. Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. England: Oxford University Press, 2017.

LIN, Patrick. *The Ethics of Autonomous Cars*. [s.L]: The Atlantic, 2013.

Machine Ethics: Creating an Ethical Intelligent Agent. Michael Anderson and Susan Leigh Anderson in AI Magazine, [s.L], v. 28, No. 4, pages 15–26; Winter 2007.

MUJICA, Fernando. Scalable electronics driving autonomous vehicle technologies. Ph.D., Texas Instruments Incorporated. 2014.

NEVEJANS, Nathalie. *European civil law rules in robotics*. Study for the European Parliament. [s.l.]: Publications Office, 2016.

NYHOLM, S. *et al*. The Ethics of Accident – Algorithms for Self-Driving Cars: an Applied Trolley Problem?. 2016. Disponível em: https://doi.org/10.1007/s10677-016-9745-2. Acesso em: 25 maio 2021.

RACHELS. *Os Elementos da Filosofia da Moral*. Barueri, São Paulo: Manole, 2006.

REINA, Giulio *et a*l. *Radar Sensing for Intelligent Vehicles in Urban Environments*. [s.L], 2015.

ROBERTS, Adrienne. Uber Crash Highlights Growing Safety Concern: Pedestrians. March 2018. Disponível em*:* www.wsj.com/articles/uber-crashhighlights-growing-safety-concern-pedestrians-1521810000?mg=prod/accounts-wsj. Acesso em: 25 maio 2021.

ROESER, S *et al.* Handbook of risk theory. Dordrecht: Springer Science, 2012.

SANDEL, Michael. *Justiça:* O que é fazer a coisa certa. Rio de Janeiro: Civilização Brasileira, 2016.

SHARIFF, A, *et al. Psychological roadblocks to the adoption of self-driving vehicles*. [s.L]. 2017.

SHARIFF, A., BONNEFON, J., RAHWAN, I. The social dilema of autonomous vehicles. *Science*, v. 352, 2016.

SILVEIRA, Matheus de Mesquisa. Tugendhat e o conceito de Moral. *Revista Eletrônica PUCRS*, Rio Grande do Sul, 2012.

SIO, Fillipo Santoni de. *Ethics and Self-driving Cars*: A White Paper on Responsible Innovation in Automated Driving Systems. 2016.

TAILLE, Yves de La. Moral e Ética: Uma leitura Psicológica. *Psicologia: Teoria e Pesquisa.* Brasília, v.26, n. especial, p. 105-114, 2010.

THOMPSON, Judith. *Killing, Letting die, and the Trolley Problem*. Oxford: University Press. 1976.

THOMPSON, Judith. The trolley problem. *The Yale Law Journal*, n. 94 , p. 1395-1415, 1985.

THOMPSON, Judith. *Turning the trolley*. [s.L]: Philosophy Public Affairs, 2008.

WEISS, Raquel. A Teoria Moral de Émile Durkheim. XIII Congresso da Sociedade Brasileira de Sociologia, Pernambuco, 2007.

WHAT is Machine Learning? A Definition. May 2020. Disponível em: https://www.expert.ai/blog/machinelearningdefinition/#:~:text=Machine%20learning%20is%20an%20application,it%20to%20learn%20for%20themselves. Acesso em: 25 maio 2021.