ethic@
Revista Internacional de Filosofia da Moral

Artigo

ON RAWLSIAN ALGORITHMS FOR AUTONOMOUS CARS

Sobre algoritmos rawlsianos para carros autônomos

Charles Feldhaus
Universidade Estadual de Londrina
Londrina, PR, Brasil
charlesfeldhaus@yahoo.com.br
https://orcid.org/0000-0001-6889-0239

A lista completa com informações dos autores está no final do artigo

Abstract: This study reconstructs some of the central aspects of Derek Leben's proposal to apply John Rawls's theory of *justice as fairness* to the development of ethical algorithms for AI-guided vehicles, as well as Keeling's criticisms that the proposal contains inaccuracies in its interpretation of Rawls's thought. It further highlights additional points where Leben's interpretation appears mistaken. Although this paper does not oppose the underlying idea itself—namely, that the application of Rawlsian algorithms to autonomous vehicles is an intellectually promising enterprise with concrete practical implications—it aims to clarify and refine the philosophical foundations of such an approach.
Keywords: artificial intelligence; maximin; justice; autonomous vehicles.

Resumo: Este estudo reconstrói alguns aspectos centrais da proposta de aplicação da justiça como equidade de John Rawls no desenvolvimento de algoritmos éticos para carros guiados com inteligência artificial de Derek Leben e algumas críticas por Keeling que a proposta contém imprecisões no que diz respeito à interpretação do pensamento rawlsiano e adiciona alguns aspectos em que a interpretação é equivocada, embora este estudo não seja contrário a ideia de pano de fundo em si mesma, a saber, de que a aplicação de algoritmos rawlsianos aos carros guiados com inteligência artificial é um empreendimento intelectual com consequências concretas promissor e digno de ser desenvolvido ainda mais.
Palavras-chave: inteligência artificial; maximin; justiça; carros autônomos.

Introduction

Technological progress has significantly transformed human life, and some even argue that the human being is an eminently technological species — that the use of

technology constitutes one of the fundamental characteristics of humankind's capacity to alter its ecological niche. In other words, unlike many other species, human beings modify their surroundings and reduce the impact of environmental conditions on their capacity for survival.

This does not, however, imply that all technological advances are necessarily positive with respect to the survival of our species. Some developments may adversely affect our environment and even place humanity at risk of extinction — as in the case of weapons of mass destruction (especially nuclear and biological weapons, though not exclusively).

One of the most balanced positions regarding scientific and technological progress, one might say, is that defended by Nicholas Agar (2015). According to Agar, technologies should be assessed according to their relationship to human well-being. This approach avoids both an extremely negative view of technology (what might be termed *radical pessimism*) and an excessively positive one (*radical optimism*).

The first position tends to attribute to technology much of the responsibility for humanity's current problems; the second, conversely, tends to see in technology the sole means of solving all human ills. The intermediate position — which Agar calls *sceptical optimism* — recognises that technologies promoting human well-being, both individual and collective, may be ethically and politically developed without restriction. Conversely, any technology that negatively affects humanity's well-being or serves the interests of a few at the expense of widespread harm should be ethically, politically, and legally constrained in its development.

With this in mind, the present study seeks to evaluate Derek Leben's approach to applying the Rawlsian theory of *justice as fairness* to autonomous vehicles guided by artificial intelligence. Leben's project rests on a reinterpretation of John Rawls's ethical model as a framework for developing moral algorithms capable of orienting decision-making in imminent accident scenarios involving self-driving cars.

Leben develops ethical algorithms inspired by Rawls's theory of justice, or at least by certain aspects of *justice as fairness*, particularly an interpretation of the maximin rule as applied to AI-guided vehicles. He first addresses this issue in a concise form focused solely on autonomous cars in the article *"A Rawlsian Algorithm for Autonomous Vehicles"* (2017) and later in a more comprehensive and sophisticated version in his book *Ethics for Robots: How to Design a Moral Algorithm* (2018). In the latter work, he expands his analysis beyond autonomous cars to include applications of artificial intelligence in both the medical and military fields.

In his earlier article, Leben's focus lies exclusively on self-driving cars from the perspective of those developing algorithms to guide ethical decisions in the face of imminent accidents. Much like Rawls's endeavour in *A Theory of Justice*, Leben aims to construct ethical algorithms as an alternative to utilitarian ones. His proposal situates the problem in relation to scenarios reminiscent of the trolley problem, operating under the assumption that one must choose between utilitarian ethics and a modified version of Rawls's maximin rule. Whereas Rawls used primary social goods as a metric of comparison, Leben employs the perspectives or survival probabilities of the self-interested parties involved in an imminent car accident — thus not morally motivated, as in Rawls's original position, but nonetheless responsible for the decision-making process.

In his later work, Leben's focus becomes broader, encompassing not only AI-driven vehicles but also other applications of artificial intelligence in healthcare and military contexts. While the range of technologies to which his model could be applied is extensive, this study will confine itself to the discussion of autonomous vehicles.

Leben begins from the premise that ethics concerns the solution of behavioural coordination problems. Drawing on insights from evolutionary biology, he views morality as an adaptive mechanism developed to resolve issues of cooperation among self-interested human beings. Consequently, the choice of the most appropriate ethical theory for designing decision-making algorithms in robots depends on which theory best addresses the problem of cooperation. Presented in this way, Leben argues, the optimal solution is a form of contractualism inspired by Rawls's justice as fairness.

I. Leben and the rawlsian algorithm applied to autonomous cars

In 2017, Derek Leben published an article in which he applies John Rawls's theory of *justice as fairness* as a moral framework for developing algorithms designed to guide AI-driven vehicles when faced with ethical dilemmas akin to the so-called *trolley problem*—that is, scenarios in which it is necessary to choose among alternative courses of action, each entailing harmful consequences for different individuals.

It is worth noting that some authors question the centrality of such dilemmas to the ethics of AI-guided vehicles, arguing that the task of philosophers should be limited to determining *when to brake*, rather than engaging in complex calculations concerning *who should be saved*. Nevertheless, a number of researchers have sought to apply major normative ethical theories—such as Kantian deontology, utilitarianism, and virtue ethics—to moral questions involving autonomous vehicles. Others, by contrast, have turned to

theories of justice, a form of ethical contractualism, most notably Rawls's *justice as fairness*.

As is well known, in *A Theory of Justice*, Rawls develops a conception of justice that applies to a society of human beings conceived as a cooperative enterprise into which individuals enter only by birth and leave only by death. In other words, although Rawls drew inspiration from the classical contractualist tradition of Thomas Hobbes, John Locke, Jean-Jacques Rousseau, and Immanuel Kant, the hypothetical choice situation of the original position does not concern an agreement to enter civil society, but rather an agreement among individuals already living in society who nonetheless disagree about how the benefits of social cooperation should be distributed.

In formulating his conception of justice for AI-guided vehicles, Derek Leben begins from a similar presupposition. He understands ethics, from an evolutionary perspective, as a social integration mechanism designed to resolve problems of cooperation among self-interested human beings. Instead of distributing the resources generated by social cooperation, however, a Rawlsian contractualist ethics for robots seeks to identify the proper course of action in ethical dilemmas comparable to *trolley problems*. In situations involving imminent car accidents, and assuming that *ethical robots* could surpass human decision-making in such contexts, Leben argues that the normative theory guiding the work of software engineers developing ethical decision algorithms should not be modelled on human moral reasoning, given that human judgement is inevitably subject to biases and personal preferences.

Leben contends that the appropriate algorithmic framework for ethical robots in the case of AI-guided cars should be a version of contractualism, specifically one that employs the maximin rule—that is, the principle of maximising the position of the least advantaged—in allocating certain types of resources. These resources, analogous to Rawls's primary goods in *justice as fairness*, refer to things that any rational person would desire: bodily integrity, health, survival, and freedom from serious injury or lifelong disability, among others. Leben even suggests that one may imagine a continuum whose worst outcome is the death of one or more individuals in an imminent collision, while the opposite extreme corresponds to escaping entirely unscathed.

The basic idea, therefore, is that when comparing alternative actions in the face of an impending traffic accident, one should first avoid choosing any alternative with a high likelihood of fatal victims. Just as Rawls's theory simplifies interpersonal comparisons by focusing on a socially relevant standpoint—the position of the least advantaged—so too, in the case of accidents involving AI-guided cars, priority should be given to those in the most

disadvantaged position: the individuals most at risk of death.

When the risk of death involves more than one person, other factors must be considered—such as the probability of death or the extent of injuries sustained by others involved. To address this complexity, where multiple fatalities may occur under different alternative actions, Leben introduces an auxiliary rule to the maximin rule, which he calls the leximin rule.

In Rawls's *A Theory of Justice*, the idea of lexical order refers to the hierarchy of priority between the two principles of justice—liberty and equality—and between the two components of the second principle—the difference principle and the principle of fair equality of opportunity. The lexical order prohibits trading off basic liberties for improved economic conditions. In Leben's framework, however, the lexical structure applies to situations of *tie* between different action alternatives in an imminent accident, with regard to the position of the least advantaged individuals.

Suppose that, across several possible actions, the probability of fatalities is identical. How should one then decide among them? Leben argues that, in such cases, we should also consider the position of the next least advantaged individual. If, for example, two alternatives present equal fatality risks, but in one of them the next least advantaged person would suffer severe injuries resulting in lifelong disability, whereas in another they would sustain only minor injuries, then the latter alternative—according to the leximin rule—ought to be preferred.

II. Keeling and some objections to Leben's algorithms

In *Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles*, Geoff Keeling argues that he rejects the contractualist response to the *moral design problem* as developed by Derek Leben in his original article. As Keeling explains, Leben's formulation of this response would require an independent argument apart from Rawls's theory of *justice as fairness*, since, as it stands, it cannot be defended in Rawlsian terms (Keeling, 2017, pp. 1–2). Keeling presents objections or challenges that he believes Leben must address if his contractualist algorithm for autonomous vehicles is to be justified independently of Rawls's original theory.

A) The use of the veil of ignorance is not justified in the context of decision-making concerning AI-guided vehicles, since, according to Rawls's own framework, such situations correspond to the fourth stage—in which the deprivation of information

has been entirely lifted (Keeling, 2017, p. 7).

B) Furthermore, the maximin rule is not a principle that Rawls regarded as universally or unconditionally applicable to all contexts and questions. Within Rawls's theory, it applies specifically to the original position, and only under certain well-defined conditions. Keeling argues that these necessary conditions are not met in the cases to which Leben seeks to apply the rule (Keeling, 2017, pp. 8–9).

C) Finally, Keeling presents several challenges which, in his view, demonstrate that Leben's Rawlsian algorithm could lead to counter-intuitive results; he even suggests an alternative algorithm to the one proposed by Leben (Keeling, 2017, pp. 9–12).

I am not entirely convinced that Keeling is correct on all the points he raises, since I believe that even though Rawls himself did not apply the veil of ignorance or the maximin rule beyond the specific context of *A Theory of Justice*, the extension of Rawls's theory has often yielded insightful and fruitful assessments of various questions of justice in diverse contexts. Nonetheless, it must be acknowledged that there are several inaccuracies in the manner in which Leben employs Rawls's theory in his account of ethical robots.

Furthermore, Keeling notes that the fact that Leben's algorithm evaluates possible actions in terms of survival probabilities may lead to counter-intuitive results, whereby in some scenarios the algorithm would select outcomes to which the passenger could not rationally consent (Keeling, 2017, p. 11). This difficulty, Keeling contends, arises from Leben's use of the maximin rule—giving priority to the least advantaged party, in this case the person most at risk in the accident—twice in his decision procedure, a move Leben himself calls the leximin rule, employed in the event of a tie between alternatives.

Keeling's second challenge states that "the maximin rule gives undue weight to the moral claims of the worst off" (Keeling, 2017, p. 11). He illustrates this through a case in which a car accident involves the potential for lifelong debilitating injuries to pedestrians. Leben's algorithm, according to Keeling, would prioritise avoiding the death of the passenger over preventing serious, lifelong injuries to several pedestrians—regardless of the number of victims involved. This, Keeling argues, seems counter-intuitive, particularly as the number of victims suffering such injuries increases relative to the single passenger who might die.

In defence of his Rawlsian algorithm, Leben replies that such scenarios—where a single fatality is weighed against numerous lifelong injuries—are highly improbable.

Moreover, he insists that if he personally were part of such a situation, he would *always* prefer to be one of the pedestrians sustaining lifelong injuries rather than the passenger who dies; in other words, he would always prefer life, regardless of the conditions in which it is lived. Yet, as is well known, this is not a universal human preference. Many individuals—particularly those facing severe suffering or loss of quality of life—may choose euthanasia, assisted death, or the withdrawal of treatment.

To illustrate with a hypothetical example: imagine that one of the accident's victims is an elite Olympic athlete who, as a result of the crash, would never again be able to compete and would spend the rest of their life bedridden or in a wheelchair. Would such a person necessarily prefer life to death in that scenario? It is difficult to say *a priori*, but Leben's argument presupposes a preference for life that cannot be taken as unanimous.

Moreover, Keeling's critique also touches upon the problem of involuntary sacrifice—the imposition of severe harm on pedestrians *without their consent* and without considering their own life preferences. Yet, it must be noted that Keeling's own position is not without controversy, since it too presupposes contested assumptions about rational consent and moral weight.

A further problematic aspect of Leben's ethical algorithm concerns his interpretation of Rawls's thought. Although Keeling (2017, p. 2) agrees with Leben's general contractualist strategy for designing ethical algorithms inspired by Rawlsian moral theory, he regards Leben's reading of Rawls as incomplete and, in some respects, inaccurate. This study aims to show that there are additional elements—beyond those identified by Keeling—that Leben also overlooks or misconstrues.

Keeling (2017) critically evaluates Leben's proposal and maintains that Leben misreads Rawls's conception of *justice as fairness* (Keeling, 2017, p. 6). He argues that several aspects of Leben's reasoning cannot be justified in Rawlsian terms. Nonetheless, Keeling (2017, p. 5) explicitly acknowledges that *"Leben and I agree more than we disagree"*—that is, he supports Leben's broader strategy of addressing the ethical questions raised by artificial intelligence in autonomous vehicles and considers a contractualist framework a promising avenue.

According to Keeling, however, Leben misunderstands Rawls's motivation for introducing the veil of ignorance, a misunderstanding he calls at least imprecise, if not outright mistaken (Keeling, 2017, p. 8). For Keeling, Rawls does not claim in *A Theory of Justice* that the maximin rule is universally applicable in all contexts; rather, it is plausible only within the specific context of the original position, as part of the rational choice among

principles of justice.

On this point, one must distinguish between contexts in which a theory of justice is being justified—as in the choice of the two principles under the conditions of the original position—and contexts in which that already justified theory is applied to practical cases. In the latter, additional elements of Rawls's theory come into play, such as the four-stage sequence and the gradual lifting of informational constraints. Once the theory is established, decision-making no longer relies on the maximin rule itself, but on the application of Rawls's principles of basic liberties, the difference principle, and the principle of fair equality of opportunity.

It is also important to note that following the publication of his article on Rawlsian applications to autonomous vehicles, Leben (2019) developed a more comprehensive and sophisticated conception of Rawlsian algorithms, extending beyond AI-guided cars to include other domains such as healthcare (saving lives), the military sphere (maintaining peace), and again autonomous vehicles (avoiding collisions with cars and pedestrians). He also sought to clarify the ethical and philosophical presuppositions of his framework by engaging with issues concerning the nature of morality (as a solution to coordination problems), moral psychology, the nature of contractualism (Rawlsian with consequentialist elements), and ethical machines in general.

Keeling (2017) maintains, among other points, that Leben's strategy cannot, in certain respects, be justified within a Rawlsian framework. It seems plausible that part of the resistance to Leben's proposal stems from a conceptual confusion or lack of clarity regarding how he employs Rawls's *justice as fairness* in his ethical algorithm for machines (and particularly for autonomous vehicles). Specifically, Leben appears to blur the distinction between:

(a) the role of the original position, in which the principles of justice are chosen among competing alternatives such as classical utilitarianism, average utilitarianism, perfectionism, and intuitionism; and

(b) the role of the principles of justice themselves, once selected, in guiding questions of basic justice within a well-ordered society capable of cultivating a sense of justice among its citizens.

When applying normative theories such as Rawls's *justice as fairness* to emerging technologies like artificial intelligence, it is essential to recognise Rawls's four stages of

application:

1. the stage of choosing principles under extensive informational constraint (the veil of ignorance) in the original position;

2. the stage of constitutional design, where some information is reintroduced, guided by the chosen principles;

3. the stage of ordinary legislation, with further reduction of informational deprivation, again guided by the previously chosen principles and constitution; and

4. the application stage, where full information is available, and decisions are made in light of the established principles, constitution, and laws.

In the case of the debate over applying *justice as fairness* to ethical problems involving AI-guided vehicles, the relevant context corresponds not to the first stage (the original position), but rather to the fourth—that of applying an already justified theory. Consequently, it is mistaken, as Leben does, to frame the issue as a theoretical contest between utilitarianism, perfectionism, and intuitionism for determining which ethical theory best suits accident scenarios involving autonomous cars.

In his later work on robot ethics, Leben—perhaps in response to Keeling's criticisms—situates the debate as if it concerned the original position itself, comparing Rawlsian contractualism (or a simplified version thereof) to libertarianism and utilitarianism. As I shall argue in the following section, several crucial aspects of Rawls's *justice as fairness*—absent from Leben's treatment—could, if properly incorporated, enhance the development of a genuinely Rawlsian approach to ethical algorithms, while others appear to have been misinterpreted or improperly assimilated in Leben's purportedly Rawlsian framework.It is likely that Leben should have employed the difference principle, which occupies a central role in interpersonal comparisons, rather than the maximin rule.

III. How rawlsian, in fact, is Leben's approach to robot ethics?

In the section addressing the four stages in *A Theory of Justice*, Rawls (1971, p. 196) asserts that he conceives the political process "as a machine which makes social decisions when the views of representatives and their constituents are fed into it. A citizen will regard some ways of designing this machine as more just than others." Rawls (1971, p. 196) maintains that a complete conception of justice does not merely evaluate laws and public

policies but establishes procedures for selecting which political opinions should be transformed into law. He understands that the sequence of stages produces a kind of framework that allows for “sorting out the complications” that must be addressed when applying the principles of justice.

As one moves from the stage of choosing the principles of justice to the “constitutional convention”, the systematic deprivation of information occurs, whereby “the veil of ignorance is partially lifted” (Rawls, 1971, p. 197). Decisions made at this point must take into account “the relevant general facts about their society”; in the case of decisions involving autonomous vehicles, it is necessary to consider relevant aspects concerning the likely collision. Rawls (1971, p. 199) distinguishes the stages based on the separation of the basic structure of society into two parts, which generally correspond to the competences of the two principles of justice: the first principle, concerning basic liberties, corresponds to the stage of the constitutional convention, while the competences of the second principle (the difference principle and the principle of fair equality of opportunity), concerning equality, correspond to the legislative stage.

Thus, one could argue that Leben’s approach to applying the maximin and leximin rules to cases involving autonomous vehicles, with significant deprivation of certain information, does not precisely correspond to Rawls’ recommended procedure. This lends some credence to Keeling’s assertion that Leben’s Rawlsian approach may lack an appropriate Rawlsian justification.

Another aspect of the relationship between stages and principles is the lexical or serial order that Rawls establishes between the two principles of justice, which Rawls (1971, p. 199) maintains throughout the division between stages, as the stage of the constitutional convention takes precedence over the legislative stage. In the fourth and final stage, concerning the “application of rules to particular cases” by judges and other public officials, as well as citizens’ compliance with the respective rules, “everyone has complete access to all the facts” (Rawls, 1971, p. 199).

Moreover, it is important to note that at this stage there is a transition from ideal theory to non-ideal theory—that is, from the level of full compliance theory to partial compliance theory—since the principles of justice derived from the ideal theory now need to be applied to concrete cases while taking into account all relevant facts pertaining to the case. The criterion for determining what knowledge is available at each stage is set “at each stage by what is required in order to apply these principles intelligently (…) while (…) any knowledge that is likely to give rise to bias and distortion (…) is ruled out” (Rawls, 1971, p. 200).

It is important to remember that artificial intelligence has, and will continue to have, a significant impact on the ways humans conduct various activities, and it is still premature to determine with certainty which applications of this new technology align with the promotion of human well-being, as suggested by Agar's cautious optimism. Nevertheless, like Leben, one may believe that autonomous vehicles can help reduce the probability of accidents, although perhaps not to the extent or in the specific ways Leben currently addresses.

There exist AI approaches to autonomous vehicles that place greater emphasis on cooperative interaction between the human driver and artificial intelligence, which can drastically reduce accidents caused by drowsiness, inattention, or speeding, among other factors. It is conceivable that fully autonomous vehicles may indeed reduce accidents in the future, and I consider the initiative of thinkers like Leben both legitimate and highly relevant. My contribution here is not to disparage this initiative or the contractarian approach, but rather to observe that ethical approaches must interact with legal frameworks—a factor that a Rawlsian algorithm must also take into account. However, as an approach inspired by Rawls' theory of justice, it overlooks certain important aspects of justice as fairness, along the lines already noted by Keeling, while going even further in the comparison.

As previously noted, I believe that Derek Leben's approach to applying justice as fairness to ethical dilemmas involving autonomous vehicle collisions overlooks the distinction Rawls draws between four stages. The first stage involves the selection of principles in the original position, where the parties are shielded by a dense veil of ignorance. The second stage entails members of society, or their elected representatives, choosing the political constitution of society while taking into account the content of the principles selected in the original position; consequently, they require more information, rendering the veil of ignorance less dense or more attenuated. The third stage consists of the selection or formulation of ordinary laws based on the principles chosen in the original position and in accordance with the chosen political constitution. Finally, the fourth stage concerns the application of laws based on the principles of justice, the political constitution, and ordinary legislation, without any deprivation of information. Leben's approach, insofar as it employs the normative principle of equality (the difference principle), properly belongs to this stage of Rawls's framework, rather than to the stage of choosing among rival theories in the original position; and at this point, full information—not informational deprivation—is what matters.

Another issue I would like to highlight regarding Leben's presentation of the problem relates to his background assumption that moral theories can be selected based on which

theory best addresses the problem of cooperation, as he states:

> In this book, I've argued that our intuitive moral judgments are adaptations to the evolutionary problem of cooperation, and moral theories are attempts to create generalized solutions to this problem. Contractarianism provides better solutions to this problem than moral theories like utilitarianism and libertarianism (Leben, 2018, p. 147).

The difficulty here is that morality, even if it may have emerged from an evolutionary perspective as an adaptive mechanism conferring advantages to humans in solving cooperation problems, does not thereby imply that morality cannot have subsequently developed beyond, or independently of, merely addressing problems of cooperation. If that is indeed the case, does it make sense to determine which algorithm is best for resolving ethical dilemmas based solely on the original evolutionary function of morality? And if morality has since evolved in other directions, becoming something distinct or significantly more complex than mere solutions to cooperation problems, does it still make sense to treat ethical questions solely as solutions to cooperation problems? I contend that the answer is negative.

A further concern pertains to the general hypothesis of Leben's book, based on the analogy between machines' superiority in board games such as chess and Go and machines' moral reasoning. He contends—which I do not dispute—that humans possess biases that often contaminate and impede our moral deliberation. Nevertheless, he appears to operate under a presumption I find difficult to reconcile, namely, that "Contractarianism matches most of our everyday intuitions": the notion that correspondence with our moral intuitions constitutes evidence for the strength of a moral theory. While this is not inherently mistaken, and such correspondence may serve as a useful criterion for selecting between moral theories, I believe Leben's theory still lacks an element explaining how to distinguish between intuitions whose correspondence is relevant and those that require correction.

At this juncture, I recall another central aspect of Rawls' theory of justice that Leben does not employ in his application to artificial intelligence, namely, the notion of reflective equilibrium. In the original edition of *A Theory of Justice*, §9, Rawls (1971, pp. 48–51) characterises reflective equilibrium as a process of mutual adjustment between considered judgments about justice and principles of justice. It is important to note that, were Leben's approach to incorporate these additional aspects of Rawlsian justice as fairness—beyond a modified version of the difference principle (or the maximin rule) that also considers other values such as liberty and equality of opportunity—it might possess further resources to

respond to Keeling's critique regarding the counterintuitive nature of permitting a growing number of individuals suffering life-long debilitating harm to be considered less significant than the death of a single person. The function of reflective equilibrium is precisely to mediate between the principles of justice (in the case of Leben's algorithm, the difference principle and the leximin rule, which also considers the second, third, least advantaged, and so forth) and our considered judgments about justice. In Keeling's example, if considered judgments point toward the choice of sacrificing one person to prevent life-long debilitating harm to many, reflective equilibrium would correct the application of the principles in such a case.

Another issue with Leben's approach concerns its conception of morality, as it appears to be committed to what Allan Buchanan argues in *Our Moral Fate: Evolution and the Escape from Tribalism* (2020), namely that there are two types of non-Darwinian dogmas in contemporary approaches to ethics in light of evolutionary biology. These are: 1) the tribalist dogma, which holds that human morality is inherently tribal and can only be tribal, meaning it is mistaken to attempt to develop normative ethical conceptions that do not acknowledge the fact that we are inherently tribal; and 2) the cooperation dogma, which maintains that human morality consists solely of solutions to social cooperation problems, and therefore, to resolve moral problems is nothing more than to resolve cooperation problems. Leben's approach does not rest upon the tribalist dogma, but rather upon the dogma of cooperation.

The first dogma is directly or indirectly related to the second, as both conflate what morality may have been when it emerged as a mechanism conferring evolutionary advantages upon the human species with what morality consists of today. That is, even if morality was tribal in origin and being tribal constituted a type of morality that conferred evolutionary advantages in the past, it does not follow that morality must remain tribal today or be understood solely as a solution to cooperation problems. In his book, Buchanan demonstrates that morality was predominantly tribal in the early stages of humanity, but two major expansions subsequently occurred: the first, approximately three hundred years ago, concerned the emergence of large-scale inclusive morality; the second involved the emergence of moralities inclusive not only towards humans but also towards non-human animals. Buchanan's objective is to show that evolutionary conservatives face considerable difficulty in explaining these two dogmas within their theoretical model. The two major expansions thus emerge as anomalies within the evolutionary theory of morality posited by conservative evolutionists.

When Leben seeks to select the best type of moral theory to serve as the basis for algorithms governing autonomous vehicles, and indeed for other applications of artificial intelligence in healthcare and military contexts, he relies on a reductionist conception of morality that fails to consider that morality may encompass elements beyond the resolution of cooperation problems. The very idea that morality may include humans unable to cooperate, such as people with disabilities, as well as non-human animals—as exemplified by the two major expansions—demonstrates that morality must address more issues than merely which theoretical moral model best handles cooperation problems among humans capable of cooperating.

It is perhaps here that Leben's approach has been influenced by the manner in which Rawls develops his own conception of justice as fairness in *A Theory of Justice*, insofar as Rawls frames his conception of justice within the context of ideal theory, that is, conditions of full compliance with the rules of justice, leaving questions pertaining to non-ideal theory for a later stage. One instance in which Rawls addresses non-ideal theory in his published work is in the final part of his book on international law, *The Law of Peoples*, in which he devotes attention to peoples who do not respect human rights, particularly outlaw states and societies burdened with unfavourable conditions.

One notable feature of Leben's proposal concerns its departure from a common assumption in philosophical ethical reflection. Typically, philosophers (and Rawls, who inspires Leben, is no exception in developing justice as fairness) rely on human moral intuitions as a benchmark for developing normative theories and for applying these theories to emerging technologies. However, Leben, inspired by board games such as chess, argues that machines can surpass humans in their capacity for ethical reasoning, just as they have done in board games. Accordingly, ethical algorithms should avoid relying solely on human moral reasoning as a standard for developing machine ethics, particularly because humans are prone to biases and errors in moral deliberation.

In other words, Leben (2018, p. 3) contends that the development of ethical algorithms for robots should not be guided by human moral psychology, as humans exhibit a "preference for people who are familiar or genetically related, ignoring the effects of our actions on people who are very distant, and relying on false beliefs about what kinds of actions are harmful." Examples of this include the non-negligible number of individuals who have condoned atrocities such as genocide, rape, and slavery.

Leben thus argues not only that the design of algorithms for machines should avoid the biases and distortions inherent in human moral reasoning, but also that, just as machines

have influenced the ways humans play board games today, machine-implemented morality may in turn shape our current understanding of morality. In this sense, Leben's approach to ethics for autonomous vehicle algorithms shares the assumption articulated by Ingmar Persson and Julian Savulescu (2012) in *Unfit for the Future: The Need for Moral Enhancement*, namely that we need to alter the way we currently reason morally. They argue for the necessity of moral enhancements, advocating the use of biomedical interventions as a complement to traditional means of improving human morality in order to address certain challenges (Persson & Savulescu, 2012).

Concluding remarks

As has been observed, Derek Leben undertakes an innovative and compelling project in applying certain aspects of Rawlsian justice as fairness, initially to autonomous vehicles and subsequently, beyond further applications to vehicles, to the domains of healthcare and the military. Leben particularly applies the maximin rule (yet he would need to apply some version of the difference principle.), which he designates, as Rawls did, the maximin rule (to grant the greatest benefit to those least advantaged). In the context of automobile accidents, the least advantaged is identified as the individual suffering the most harm in terms of health, bodily integrity, and ultimately, the potential loss of life resulting from the accident.

Where there is a tie in identifying the least advantaged, Leben introduces a second priority rule—which, of course, does not exist in Rawls' theory—the leximin rule, which assigns priority to the second least advantaged in cases of a tie among alternatives in the worst position. Keeling has published an article critiquing certain aspects of Leben's Rawlsian approach, including the very possibility of justifying Leben's method in Rawlsian terms.

Moreover, this study expands upon Rawlsian aspects that Leben's approach does not consider but arguably should, and it interrogates the underlying assumption that moral issues can be equated solely with solutions to cooperation problems. Even if morality may have originally emerged as a solution to cooperation problems, it is not evident that all aspects of morality today can be understood solely through this lens.

References

AGAR, Nicholas. **The sceptical optimist**: why technology isn't the answer to everything. Oxford: Oxford University Press UK, 2015.

BUCHANAN, Allan. **Our moral fate**: evolution and the escape from tribalism. Cambridge: The MIT Press, 2020.

KEELING, G. Against leben's rawlsian collision algorithm for autonomous vehicles. *In*: MÜLLER, V. (eds). Philosophy and theory of artificial intelligence: studies in applied philosophy, epistemology and rational ethics, v. 44. **Springer**, 2017. https://doi.org/10.1007/978-3-319-96448-5_29.

LEBEN, D. A Rawlsian algorithm for autonomous vehicles. **Ethics Inf Technol** v.19, p. 107–115, 2017. https://doi.org/10.1007/s10676-017-9419-3.

LEBEN, Derek. **Ethics for robots**: how to design a moral algorithm. New York: Routledge, 2018.

PERSSON, Ingmar; SAVULESCU, Julian. **Unfit for the future**: the need for moral enhancement. Oxford: Oxford University Press, 2012.

RAWLS, John. The law of peoples. **Critical Inquiry**, v. 20, n. 1, p. 36-68, 1993.

RAWLS, John. **A theory of justice**. Revised Edition. Cambridge: Harvard University Press, 1999 [1971].

DADOS DE PUBLICAÇÃO

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: C. Feldhaus
Revisão e aprovação: C. Feldhaus

AGRADECIMENTOS
Agradeço aos membros do Grupo de Trabalho do GT Teorias da Anpof que participaram da apresentação do presente estudo na forma de palestra durante o encontro do GT durante o Simpósio Internacional Principia em Florianópolis em 2025 e que indiretamente podem ter contribuído para a realização dessa versão final submetida à revista.

USO DE INTELIGÊNCIA ARTIFICIAL
O artigo foi escrito sem ajuda de nenhuma inteligência artificial na íntegra inicialmente em português. O autor somente utilizou alguma ajuda do ChatGPT (GPT-4) para revisão gramatical do texto em inglês e sugestões de frases no momento da tradução do português ao inglês do artigo. Todo o conteúdo argumentativo foi desenvolvido integralmente pelo autor sem ajuda da inteligência artificial.

FINANCIAMENTO
Agradeço à Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná através do Edital PROPPG 05/2025 - APOIO INSTITUCIONAL À PARTICIPAÇÃO EM EVENTOS TÉCNICO-CIENTÍFICOS da Universidade Estadual de

Londrina pelo apoio financeiro para a minha participação no 14th Principia International Symposium que motivou a confecção do presente artigo.

CONFLITO DE INTERESSES

A pessoa autora declarou não haver interesses conflitantes.

DISPONIBILIDADE DE DADOS DE PESQUISA E OUTROS MATERIAIS

Os dados completos foram publicados no próprio artigo. Todo o conjunto de dados que dá suporte aos resultados deste estudo está incluído no corpo do artigo.

LICENÇA DE USO

PUBLISHER

Universidade Federal de Santa Catarina. As ideias expressadas neste artigo são de responsabilidade das pessoas autoras, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES

Darlei Dall'Agnol
Vitor Sommavilla de Souza Barros
Odair Camati

HISTÓRICO

Recebido em: 31-10-2025
Aprovado em: 28-03-2026
Publicado em: 30-04-2026