第５３回　日本科学哲学会

ワークショップ「社会の中の道徳的ジレンマ」

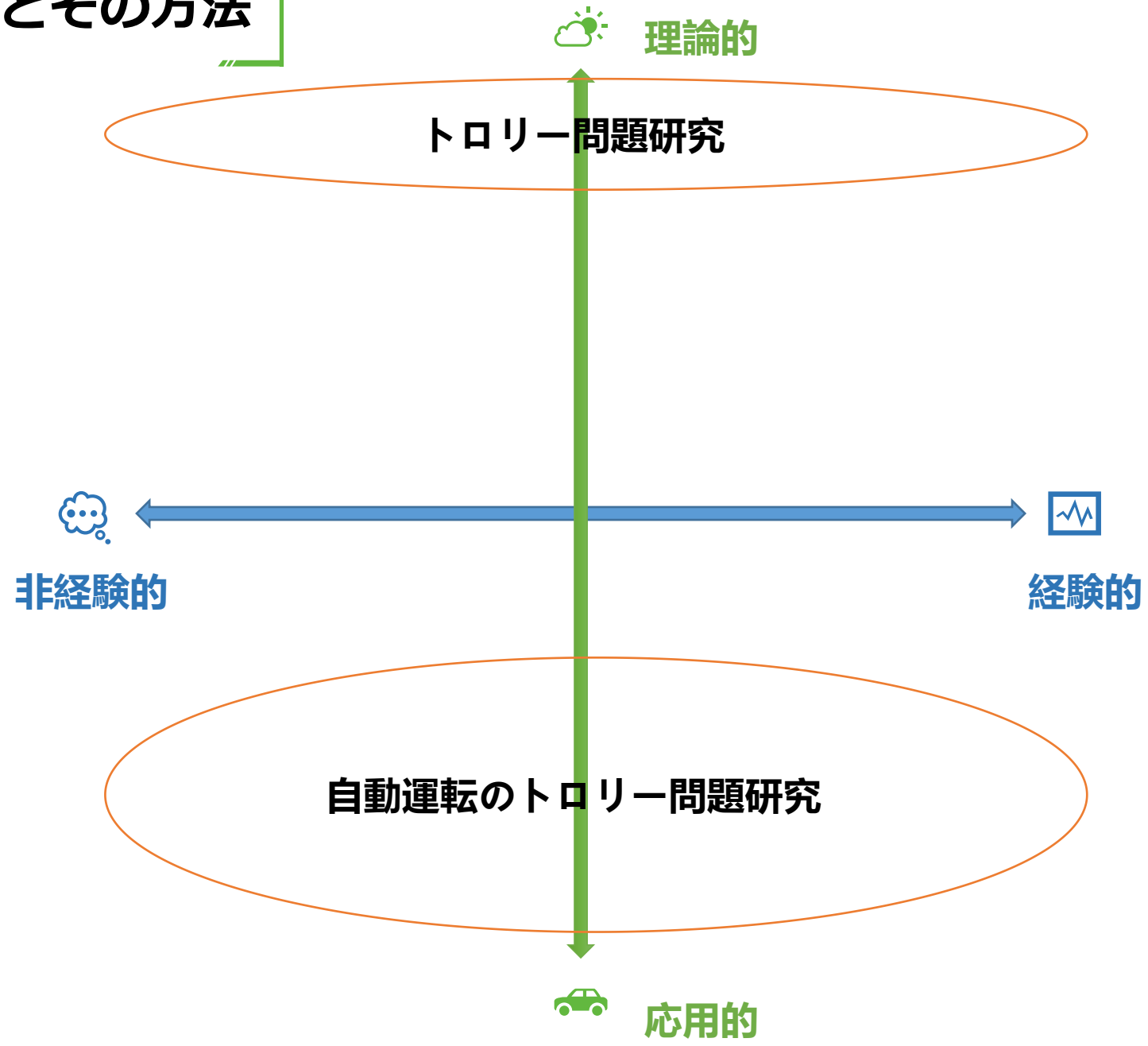
自動運転のトロリー問題から考える応用倫理学の方法論

笠木雅史（名古屋大学）

本研究は、トヨタ財団特定課題「先端技術と共創する新たな人間社会」　：社会的意思決定を行うAIの要件の助成を受けています）

2020年10月11日

倫理学とその方法



内容

1. 規範倫理学におけるトロリー問題
2. 実験哲学におけるトロリー問題
3. 自動運転のトロリー問題
4. 応用倫理学の方法論

1

規範倫理学におけるトロリー問題

Philippa Foot (1968)

事例(1) そのままでは生存の見込みがない５人の患者がいる。医師がその５人に移植するために、１人の健康な人の臓器を取り分ければ、５人の生命は助かるが、１人の生命は失われる

事例(2) 暴走するトロリーの路線上には追突必死の作業員５人がいる。運転手がその５人を救うために、１人の作業員のいる路線へとトロリーを車線変更すれば、５人の生命は助かるが、１人の生命は失われる

Footの判断：事例(1)では医師の移植行為は道徳的に許容不可能だが、事例(2)では運転手の路線変更行為は許容可能

その根拠：事例(1)では１人の殺人（作為）と５人を死ぬにまかせる（不作為）が問題になっているのに対し、事例(2)では１人の殺人と５人の殺人が問題になっている

Judith J. Thomson (1976), (1985)

事例(3) 暴走するトロリーの路線上には追突必死の作業員５人がいる。線路脇の路線変更スイッチの前には第三者がいるが、５人を救うために、１人の作業員のいる路線へとこの第三者がトロリーを車線変更させれば、５人の生命は助かるが、１人の生命は失われる

事例(4) 暴走するトロリーの路線上には追突必死の作業員５人がいる。線路の跨線橋上には第三者と１人の体格の大きな人物がいる。その体格でトロリーの進行を止めるために、第三者がこの人物を線路に突き落とせば、５人の生命は助かるが、１人の生命は失われる

事例(5) 暴走するトロリーの路線上には追突必死の作業員５人がおり、その路線から分かれる他の路線には１人の体格の大きな人物がいる。線路脇の路線変更スイッチの前には第三者がいる。その体格でトロリーの進行を止めるために、この人物がいる路線へと第三者がトロリーを車線変更させれば、５人の生命は助かるが、１人の生命は失われる。車線変更先の路線は、大柄の人物の向こう側でもとの路線に再合流し、この人物がトロリーを止めなければ、再び作業員５人に向う

Thomsonの判断：事例(5)では事例(3)と同様に、第三者の行為は道徳的に許容可能だが、事例(5)では事例(3)とは異なり、第三者は体格の大きい人物の死を意図している⇒殺人と死ぬに任せるの区別では、この点を説明できない

※Thomsonが「トロリー問題」と呼んだのは、事例(1)と事例(2)の許容可能性についての相違を説明することであるが、第三者観点からの事例(3)と事例(4)の相違も同じ名前で呼んでしまった

道徳的ジレンマへの一般化

道徳的ジレンマ状況

(a) 特定の行為者が何を行なうのかを選択し、実行できる

(b) その行為者にとって可能な行為の選択肢は限られている

(c) 複数の人やグループが、選択される行為の影響を受ける

(d) 選択されるどの行為によっても、ある人やグループに重大な損害が生じる

トロリー問題の解決

(a) 特定の事例間、あるいは道徳的ジレンマ状況一般において、特定の行為（選択肢）が道徳的に許容可能・不可能かどうかを正当化する道徳的要因を特定する

(b) 道徳的に許容可能・不可能かどうかを正当化する道徳的原理を特定する

(a') 特定の事例間、あるいは道徳的ジレンマ状況一般において、特定の行為（選択肢）が道徳的に許容可能・不可能かどうかを説明する道徳的要因を特定する

(b') 道徳的に許容可能・不可能かどうかを説明する道徳的原理を特定する

反照的均衡法

反照的均衡法

様々な現実的事例と仮想的事例における道徳的許容可能性・不可能性についての判断を調査し、そうした判断と最大限に一致する原理を抽出すること

トロリー問題への反照的均衡法への批判

(1) 現実にはありそうにない事例を考察対象としている⇐様相懐疑論

(2) あまりにも抽象的に事例が記述されており、判断材料が少なすぎる⇐Swansea School

(3) 本来深刻な問題を戯画化された仕方で扱っている⇐Anscombe

(4) 道徳的要因の普遍化可能性を前提にしている⇐個別主義

(5) 事例の表面的で非道徳的な特徴に判断は左右される⇐実験哲学

(6) 判断の真理と外延的に一致する要因を特定するだけであり、説明を与えない⇐Kegan

説明性要求

結局のところ、私たちが受容可能な道徳的原理に求めるのは、それが個別事例についての直観に一致するということだけではない。それに加え、私たちは原理が訴える様々な要因がなぜ道徳的に問題になるのかを理解したいと求めるのである。(Kegan 2015: 153)

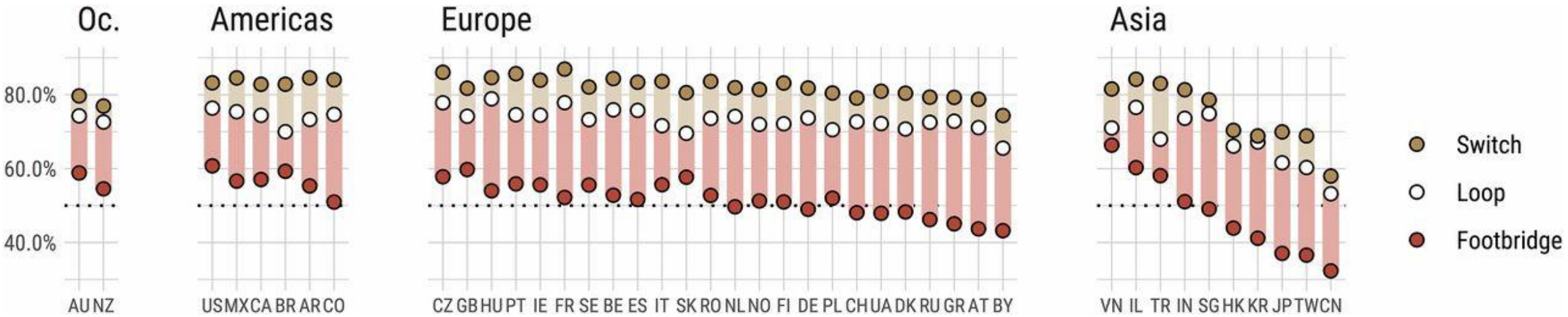
2

実験哲学におけるトロリー問題

判断の心的要因の研究(1)

トロリー事例についての心理学的・神経科学的研究

1990年代から、心理学において質問紙調査によってトロリー事例を含む道徳的ジレンマ状況についての判断の研究が始まる。さらにその後脳機能イメージを用いた研究も加わり、現在は非常に大規模になっている



Awad et al. (2020)
42ヶ国、10の異なる言語により、合計7万人の参加者に対して実施された調査
事例(3)を許容可能と判断する人数がもっとも多く（平均81%）、次いで事例(5)（平均72%）、事例(4)（平均51%）という順番となったが、文化差がある

判断の相違とその要因(2)

トロリー事例内の判断の心理的要因

(a) 行為の意図の有無やその直接性

(b) 行為によって引き起こされる損害の種類

(c) 各行為が引き起こす損害と利益の相違

(d) 各行為によって損害が引き起こされる確率

(f) 損害が引き起こされる手段の物理的な直接性

トロリー事例外の判断の心理的要因

(a) 判断者の社会的属性（文化、性別など）

(b) 事例の提示順序

(c) 事例の記述の仕方（助けるvs殺すなど）

(d) 判断する視点（当事者vs観察者）

事例(6) 暴走するトロリーの路線上には追突必死の作業員５人がいる。線路の跨線橋上には落下装置と１人の体格の大きな人物がおり、そこから離れた地点に第三者がいる。その体格でトロリーの進行を止めるために、第三者が落下装置の遠隔スイッチを押し、この体格の大きな人が線路に落ちれば、５人の生命は助かるが、１人の生命は失われる

Greene et al. (2009)
事例(4)では突き落としは許容不可能だと判断されることが多いが、事例(6)ではスイッチを押すことは許容可能と判断されることが多い⇒反照的均衡法への批判

3

自動運転のトロリー問題

自動運転の倫理学

自動運転についての３つの道徳的問題

(1) 道徳的ジレンマ状況でのどのような操作が道徳的に許容可能なのか

(2) 自動運転によって生じた損害に対して誰がどれだけ責任を負うべきなのか

(3) 自動運転が広く社会実装されることが、どのような道徳的影響をもつか

(1)の問題は、一般化されたトロリー問題と類似性が高いため、トロリー問題と同一視されたり、トロリー問題についての議論と接続されたりすることが多い。ドイツ連邦交通・デジタルインフラ省が2017年に発行したガイドライン自動運転技術の開発、実装、使用ガイドラインは、トロリー問題に言及

(1)の問題の研究も、道徳的ジレンマ状況での選択肢の許容可能性・不可能性の正当化要因についての研究と、人々が道徳的許容可能性・不可能性についての判断を行う際の心理的要因についての研究に分かれる

心理的要因についての研究(1)：文化差

モラル・マシーン・プロジェクト(Awad 2018)

自動運転車が直面しうる様々な道徳的ジレンマ状況について、世界中から４万近い判断が集計された

統制された条件：

(1) 人間を助けるか、動物を助けるか
(2) 車線変更するか、しないか
(3) 搭乗者を助けるか、歩行者を助けるか
(4) より多くの生命を助けるか、より少ない生命を助けるか
(5) 男性を助けるか、女性を助けるか
(6) 若者を助けるか、老人を助けるか
(7) 法律を守りながら横断している歩行者を助けるか、法律を守らないで横断している歩行者を助けるか
(8) 標準的な体型の人を助けるか、大柄な体型の人を助けるか
(9) 社会的地位の高い人を助けるか、社会的地位の低い人を助けるか

(a) 個人主義的文化圏の参加者の方が共同体主義的文化圏の参加者よりも、より多くの生命を助け、若者を助ける選択肢を選ぶ傾向性が強かった

(b) 経済指標や法の支配指標が高い国の参加者ほど、法律を守らずに横断している人を助けない選択肢を選ぶ傾向性が強かった

(c) ジニ係数が高く不平等が顕著な国の参加者ほど、問題となる社会的地位の高さによって判断が異なる傾向があった

(d) 女性を助けると判断する傾向性はほとんどの国で強いが、女性の出生率と平均寿命が男性より高い国の参加者ほど、その傾向性は強かった

心理的要因についての研究(2)：当事者・観察者バイアス

Bonnefon, Shariff, & Rahwan (2016)

自動運転車の搭乗者vs観察者

道徳的ジレンマ状況で搭乗者が犠牲になるとしても、自動運転車がより多くの生命を救うことを選択するのは許容可能だと観察者として判断する人は多い。

しかし、そうした人たちでも、自動運転車を購入するならば、搭乗者を助ける選択をする自動運転車を購入したいと回答した

Kallioinen et al. (2019)

自動運転車の搭乗者vs手動運車の運転者vs歩行者

車線変更すれば自分に損害が生じるとき、自動運転の場合、どの視点でも、損害が及ぶ人数がもっとも少なくなる選択肢を選ぶのが好ましいとされたのに対し、手動運転の場合は、視点により判断が異なる傾向性が見いだされた

歩行者の視点からは、一人でも歩行者が犠牲になるならば、車線変更し運転手が犠牲になる方がほうが好ましいと判断された。他方、運転手の視点からは、２人以上の歩行者、観察者の視点からは、４人以上の歩行者が犠牲になるときのみ、車線変更し運転手が犠牲になるほうが好ましいと判断された

正当化要因についての研究：背景

自動運転技術の設計・実装において、トロリー問題に注目が集まったのは、自動運転車に搭載されるAIアルゴリズムにどのように道徳的要素を組み込むかという文脈

トロリー事例のさらなる一般化

このため、道徳的ジレンマ状況一般での自動運転車の選択が問題とされる

しかし、さらに広く通常の走行条件下の条件も含める形で道徳的ジレンマを理解する者もいる。例えば、通常のカーブ走行時であっても、歩行者が急に飛び出してくる可能性を考えると、どれくらいの速度でカーブを曲がるのが許容可能かという問題は生じる

正当化要因についての研究

損害の総量がもっとも少なくする功利主義的なアルゴリズム(Awad et al. 2018)

自動運転車の行動選択では、搭乗者や歩行者、自動運転車の所有者、その開発者、そして交通法の制定者などに損害が生じる可能性がある。どれか一つの人物に対する損害を重視し、それを軽減するようなアルゴリズムは、他の人物に対する損害を深刻化する

また、調査で示されたように、どのような視点を採用するのかで、人々の判断は変化する。特定の視点の選好を反映するアルゴリズムは、結果として全体的な交通安全の上昇や事故数の抑制といった、自動運転車を導入する最大の理由と衝突する

マキシミン・ルールに基づく社会契約主義的な自動運転のアルゴリズム(Leben 2017)

利益を単純に、身体的な損害の少なさ、つまり生存確率と同一視した上で、マキシミン・ルールを自動運転に適用する。各選択肢においてもっとも生存確率が低い人物や集団の生存確率を比較し、相対的にそれがもっとも高い選択肢が最善のものとなる

ハンドルをきらずブレーキだけをかける(Davnall 2020)

自動運転車に搭載されるセンサーからの情報はつねに限定的であり、また自動車の挙動も路面の状態、天候、またブレーキの性能という現実の条件によって左右される。こうした点を認めるならば、自動運転車が直面する現実の条件下では、特定の事故を回避するために、ブレーキをかけながらハンドルをきるよりも、ハンドルをきらずにブレーキだけをかける方が、危険な事故が起こる確率は低い

4

応用倫理学の方法論

自動運転のトロリー問題への批判

(1)	自動運転車が適切に動作している環境下では、大きな被害につながる事故は起きず、起こる場合には何らかの誤作動により選択肢を選ぶことができないか、選んでもその通りの動作ができない環境であるため、道徳的ジレンマ状況に自動運転車が直面する確率は極めて低い(Himmelreich 2018)
(2)	仮想的な道徳的ジレンマ状況は、利害関係や責任・保障の義務などを捨象しており、さらにどの選択肢がどの結果につながるのかも確定しているため、事故を含む結果が不確実な現実の条件とあまりにも異なりすぎている(Goodall 2016; Nyholm & Smids 2016; Mirnig & Meschtscherjakov 2019)
(3)	自動運転のためのAI開発は、規則ベースではなく深層学習により行われるため、規範倫理学におけるトロリー問題を解決する道徳的原理の探求との関連性が薄い(Gurney 2016; Himmelreich 2018)
(4)	自動運転車の動作についての道徳的に正しい選択肢を探すのではなく、社会的に合意可能な選択肢をすり合わせていく実践的解決の方が現実的な目標となる(Himmelreich 2018)
(5)	何らかの損害が生じる道徳的ジレンマ状況に陥ることを防止するための技術、制度のほうが重要である(Mirnig & Meschtscherjakov 2019)

現実の事例と仮想的事例の相違

現実の事例と仮想的なトロリー事例での選択は、多くの点で異なっている

現実の事例	仮想的事例
具体性（細部の確定性）	抽象性（細部の不確定性）
不確実性のもとでの選択	確実性のもとでの選択
視点の非限定性	視点の限定性
様々な道徳的・非道徳的要因の関与	限定的な道徳的・非道徳的要因の関与
長期的・短期的時間要因が関与	短期的時間要因のみが関与

どちらの判断の場合も、具体性、関連要因の多さ、不確実性がトレードオフの関係にある。どれかを絞ることが分析のためには必要となることもある

仮想的事例では、多くの道徳的要因や時間的要因が捨象されているため、それらは別個に考察されなければならない。仮想的なトロリー事例における道徳的許容可能性は、pro tantoなもの。判断の心理的要因、事例内の要因を考察すれば、阻却されることは十分ある

道徳的説明への方法論的洗練

単なる「許容可能性」ではなく、「道徳的許容可能性」を説明する道徳的要因を特定することが重要

事例についての判断を行う際は、自分の判断が道徳的に重要なのかを確かめること。道徳的重要性についての判断は、許容可能性についての判断が道徳的要因をトラックしているかを確かめるのに有効(Andow 2018)

反照的均衡法は、正当化のための方法であるため、「道徳的説明」についての方法論によって補完される必要がある。基本的な道徳的要因の事例に合わせた具体化が必要（例えば、Kagan (2015) は、行為の目的vs手段の対比による説明を、人の代替性vs人の従属による説明に置き換える）。この具体化は、事例ごとに異なる

その上で、他の道徳的要因、非道徳的要因の考慮を行う

参考文献

Andow, J. (2018). “Are intuitions about moral relevance susceptible to framing effects?” Review of Philosophy and Psychology, 9(1): 115-141.

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

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). “The social dilemma of autonomous vehicles.” Science, 352(6293): 1573-1576.

Davnall, R. (2020). “Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics.” Science and Engineering Ethics, 26 (1): 431-449.

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

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. (2009). “Pushing moral buttons: The interaction between personal force and intention in moral judgment.” Cognition, 111 (3): 364-371.

Goodall, N. (2016). “Away from trolley problems and toward risk management.” Applied Artificial Intelligence, 30 (8): 810-821.

Gurney, J. K. (2016) “Crashing into the unknown: an examination of crash-optimization algorithms through the two lanes of ethics and law.” Albany Law Review, 79(1): 183–267.

Himmelreich, J. (2018). “Never mind the trolley: The ethics of autonomous vehicles in mundane situations.” Ethical Theory and Moral Practice, 21(3): 669-684.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Stephan, A., Pipa, G., & König, P. (2019). “Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives.” Frontiers in Psychology, 10: 2415.

Kagan, S. (2015). “Solving the trolley problem.” In F. M. Kamm & E. Rakowski (eds.), The Trolley Problem Mysteries, Oxford University Press: 151-168.

Leben, D. (2017). “A Rawlsian algorithm for autonomous vehicles.” Ethics and Information Technology, 19 (2): 107-115, 2017.

Mirnig, A. G. & Meschtscherjakov, A. (2019) “Trolled by the trolley problem: on what matters for ethical decision making in automated vehicles.” Proc. of the 2019 CHI Conference on Human Factors in Computing Systems: 1-10.

Nyholm, S., & Smids, J. (2016). “The ethics of accident-algorithms for self-driving cars: An applied trolley problem?” Ethical Theory and Moral Practice, 19(5): 1275-1289.

Thomson, J. D. (1976). “Killing, letting die, and the trolley problem,” The Monist, 59 (2): 204-217.

Thomson, J. D. (1985). “The trolley problem,” Yale Law Journal, 94 (6): 1395-1415.