博士学位論文審査要旨

２０２１年１２月２２日

論文題目：　自動運転車に対する信頼の規定因の検討―道徳判断の一致による効果―

学位申請者：　横井　良典

審査委員：

主　査：　心理学研究科　教授　中谷内　一也

副　査：　東京大学大学院人文社会系研究科　教授　唐澤　かおり

副　査：　心理学研究科　教授　神山　貴弥

要　　旨：

近年、人工知能の導入がさまざまな領域において進められている。人工知能の判断が人の生命や健康、財産を直接左右すると知覚される領域においては、信頼がその導入の鍵を握ることになる。では、人工知能への信頼は何によって規定されるのか。本論文は価値の共有がリスク管理者への信頼を説明するという主要価値類似性モデルを基盤とし、道徳判断が主要価値を反映するという考えのもと、信頼への影響を検討している。一連の実験では、モラルジレンマシナリオでしばしば用いられるトロッコ問題のバリエーションを採用し、可能な限り犠牲者を減らそうとする功利主義的な判断と、意図的な行為によって人に犠牲を強いるべきではないという義務論的判断との対立に焦点をあてている。

第１研究では、参加者の道徳的選好と自動運転車の判断とが功利主義で一致している場合、参加者の自動運転に対する信頼が高くなることが示された。一方、義務論的判断が選好されやすいような状況では、参加者の道徳的選好と自動運転車との判断が義務論で一致しても信頼への影響はみられなかった。

そこで第２研究では、ジレンマ状況の設定や実験参加者の属性を拡張して、第１研究の結果の一般性が検討された。具体的には、自動運転に加え、治療を優先すべき患者を選定するトリアージ場面を設定し、人工知能に対する信頼が価値の共有によって説明されるのかが検証された。２つの実験結果から、功利主義・義務論に関わらず、参加者が選好する道徳判断と人工知能が下す道徳判断が一致していると信頼が高くなることが示された。

さらに第３研究でも、実験参加者の属性を拡大して第１研究の追試が行われ、功利主義だけでなく、義務論的判断が選好される場合でも、価値の一致が自動運転への信頼を説明することが明らかにされた。

一連の研究を通して、参加者の属性やシナリオの領域を超えて道徳判断一致の効果が見出され、人工知能、とくに、自動運転車への信頼に関する主要価値類似性モデルの説明力が示された。

これまで、価値共有による信頼への効果が数多く報告されてきたが、それらはリスク管理をする人間への信頼に限定されていた。それに対して、本論文は対象が人工知能であっても、リスク判断をする対象への信頼が道徳的な価値共有によって説明されることを示す初めての研究である。本研究の知見は信頼研究における理論的な発展のみならず、実務的な示唆という点においても有意義なものと評価される。

よって、本論文は、博士（心理学）（同志社大学）の学位を授与するにふさわしいものであると認められる。

総合試験結果の要旨

２０２１年１２月２２日

論 文 題 目：　自動運転車に対する信頼の規定因の検討―道徳判断の一致による効果―

学位申請者：　横井　良典

審 査 委 員：

主　査：　心理学研究科　教授　中谷内　一也

副　査：　東京大学大学院人文社会系研究科　教授　唐澤　かおり

副　査：　心理学研究科　教授　神山　貴弥

要　　旨：

上記審査委員３名は、2021 年 12 月 17 日(金)午後 0 時 20 分より 40 分間に及ぶ博士学位論文公聴会の後、午後１時５分より２時間にわたって学位申請者に対して総合試験を行った。

学位申請者は、提出した論文に関する審査委員からの専門的質疑に対して、適切な説明と応答を行い、本論文の学術的価値を証明した。また、申請者は本研究の基礎となる社会心理学および産業心理学領域について、広範な専門的知識を持ち合わせていることが確認された。さらに、引き続き実施された口頭試問による語学試験と第一著者として発表された英語論文３件についての質疑から、十分な語学力（英語）を有することも認められた。

よって、総合試験の結果は合格であると認める。

博 士 学 位 論 文 要 旨

論 文 題 目：　自動運転車に対する信頼の規定因の検討
―道徳判断の一致による効果―

氏　　名：　横井　良典

要　　旨：

近年，人工知能技術がさまざまな分野で応用されている。中でも，自動運転車においてその技術の応用が活発に進められている。自動運転車の導入によって，交通事故の減少，渋滞の緩和といったメリットが期待されている (Waldrop, 2015)。我が国においては，2025 年頃には高速道路限定ではあるが，人間の操作を介さない完全自動運転車の導入が予定されている (国土交通省，2019)。将来の動向を踏まえ，本研究では，人間の操作を介さない完全自動運転車を題材に，自動運転車への信頼の規定因を検討する。

自動運転車が導入されたとき，自動運転車を信頼できるかどうかが重要な問題となる。信頼の定義について，Lee & See (2004) は「被害を受ける可能性がある状況において，個人の目標を達成するために，機械に援助してもらおうとする態度」と定義している。交通場面においても，自動運転車に運転を任せることで，交通事故などの被害を受ける可能性が想定される。本研究では，自動運転車への信頼を，被害可能性を受け入れてでも自動運転車に運転を任せようとする態度として扱う。自動運転車によるメリットを生かすためには，そのような信頼が重要になるだろう。

信頼が重要であれば，次に問題になるのが信頼の規定因である。Earle & Cvetkovich (1995) は信頼の規定因として価値の共有を挙げ，「ある問題に対する見立て，その問題に関わる目的や目的を達成するまでのプロセスについて，自身が重要視する価値を相手も持っていると，その相手を信頼する」と述べている。リスク認知研究では，「自身が重要視する価値を相手も持っている」という認知，すなわち価値共有認知がリスク管理者への信頼を説明することが示されてきた (Siegrist, 2021)。本研究では，自動運転車と価値を共有していると思うかどうかという価値共有認知を実験的に操作し，その操作によって自動運転車への信頼が変わるのかどうかを検討する。

本研究では，価値として道徳判断を扱い，人々が望む道徳判断と自動運転車が下す道徳判断を一致させるかどうかによって価値共有認知を操作し，自動運転車への信頼が変化するのかどうかを検討する。道徳判断の中でも，トロッコ問題などのモラルジレンマシナリオでよく用いられる功利主義と義務論という 2 つの判断を扱う。トロッコ問題とは，「暴走したトロッコが 5 人の作業員に向かって走っている。このまま直進すると，この 5 人の作業員を轢いてしまう。この 5 人を救うには，レバーを引いて線路を切り替える必要がある。しかし，線路を切り替えると，その先にいる別の 1 人の作業員を轢いてしまう」というシナリオである (Foot, 1967)。このとき，線路を切り替えて 1 人を轢く判断を功利主義的な判断と呼ぶ。功利主義とは幸福の最大化，多くの人の利益を追求するといった考えである (Bentham, 1789/1967)。一方，直進して 5 人を轢く判断を義務論的な判断と呼ぶ。義務論には，人々は決められた義務を果たせなければならない，意図的に危害を加えてはならないという考えが含まれている (Kant, 1785/1976)。トロッコ問題に当てはめると，線路を切り替えるという意図的な行為によって，1 人の作業員を犠牲にするべきではないというのが義務論の考えである。本研究でも，義務論を意図的な行為によって人々を犠牲にしてはならないという考えとして扱う。Earle & Cvetkovich (1995) は，目的のようなお互いにとって重要な価値の共有が相手への信頼を決めると述べている。功利主義や義務論といった判断は，最終的に誰を守るかという目的に該当し，道徳判断は重要な価値と認識されるだろう。それゆえ

に本研究では，価値として道徳判断を扱う。

本研究の目的は，価値共有認知を実験的に操作することで，自動運転車への信頼が変化するのかどうかを検討することである。この目的に沿って，実験では道徳判断の一致 (一致条件 vs 不一致条件，参加者間要因) を操作した。道徳判断の一致について，実験参加者が望む道徳判断と自動運転車が行う道徳判断を一致させるかどうかによって，価値共有認知を操作した。実験操作のために，トロッコ問題のようなモラルジレンマシナリオを作成した。シナリオ実験は剰余変数の統制など，内的妥当性を高めるうえで有効な手段である (Schafheitle, Weibel, Meidert, & Leuffen, 2019)。本研究は，道徳判断の一致による自動運転車に対する信頼への効果を検討する初めての研究であるため，内的妥当性を高めることを重視した。

シナリオ実験では初めに，参加者自身が功利主義的な判断を望むか，義務論的な判断を望むかを測定した。参加者に判断させた後，同様の状況において，自動運転車が道徳判断を行うシナリオを参加者に読ませた。このとき一致条件であれば，自動運転車は参加者と同様の判断を行った。一方，不一致条件であれば，自動運転車は参加者と異なる判断を行った。最後に，自動運転車に対する信頼を 3 つの質問項目を用いて測定した。

まず Yokoi & Nakayachi (2021, *Human Factors* (以下，*Hum Fac*)) では，道徳判断の一致による自動運転車への信頼の効果を検討するために，大学生を対象に 3 つの実験を行った。実験 1 (N=128) では，直進して 5 人を轢くか，車線を切り替えて 1 人を轢くかというジレンマシナリオが用いられた。この場合，5 人を救う判断が功利主義，1 人を救う判断が義務論に該当する。実験 1 のシナリオでは，参加者の約 8 割が功利主義を選好した。実験の結果，参加者の選好と自動運転車の判断が功利主義で一致している方が，一致していないときよりも自動運転車への信頼が高くなることが示された。実験 2 (N=71) と実験 3 (N=196) では，直進して 5 人を轢くか，車線を切り替えて女性と赤ちゃんを轢くかというシナリオが用いられた。この場合，5 人を救う判断が功利主義，女性と赤ちゃんを救う判断が義務論に該当する。実験 2 では参加者の約 8 割が，実験 3 では約 6 割が義務論を選好していた。実験の結果，参加者の選好と自動運転車の判断が義務論で一致しようがしまいが，自動運転車への信頼はあまり変わらないことが示された。

続く Yokoi & Nakayachi (2021, *International Journal of Human-Computer Interaction* (以下，*IJHCI*)) では，道徳判断一致の効果について，その一般化可能性が検討された。実験は 2 つ実施され，実験 1 は大学生 270 名，実験 2 は大卒の一般人 605 名のデータを収集した。道徳判断一致の操作手続きは，Yokoi & Nakayachi (2021, *Hum Fac*) と同様であった。実験 1 では信頼の対象として自動運転車が取り上げられた。シナリオは，直進して 2 人を轢くか，車線を切り替えて 1 人を轢くかという内容であった。実験 2 では，参加者を大卒の一般人，信頼の対称を医療用人工知能に変更して実験を行った。中でも，治療を優先すべき患者を選定するトリアージ場面における人工知能を題材にした。シナリオの内容は，先に病院にいる 1 人の患者を優先して治療するか，後から病院に運ばれてくる 2 人の患者を治療するかというものであった。実験 1 のシナリオでは約 7 割の参加者が功利主義を選好し，実験 2 では約 8 割の参加者が義務論を選好していた。実験 1 と 2 の結果，功利主義・義務論に関わらず，参加者が選好する道徳判断と自動運転車が下す道徳判断が一致している方が，一致していないときよりも，自動運転車や医療用人工知能への信頼が高くなるという知見が得られた。参加者の属性やシナリオの領域を超えて，道徳判断一致の効果が検出されたことから，その効果の一般化可能性が示唆された。

最後に Yokoi & Nakayachi (2021, *The Japanese Journal of Experimental Social Psychology* (以下，*JESP*)) では，Yokoi & Nakayachi (2021, *Hum Fac*) の実験 2 と 3 の追試を行った。実験は大卒の一般人 609 名を対象に行われた。実験の結果，功利主義・義務論という判断のタイプに関わらず，参加者の選好する道徳判断と自動運転車が下す道徳判断が一致している方が，一致していないときよりも，自動運転車への信頼は高くなることが示された。

Yokoi & Nakayachi (2021, *Hum Fac*) の実験 1，Yokoi & Nakayachi (2021, *IJHCI*) の実験 1 と 2，

Yokoi & Nakayachi (2021, *JESP*) の実験において，道徳判断一致の主効果が検出されたので，功利主義と義務論に関わらず，参加者が望む道徳判断と自動運転車が下す道徳判断が一致していると，自動運転車への信頼が高まると結論付けられるだろう。ただし，Yokoi & Nakayachi (2021, *Hum Fac*) の実験 2 と 3 の結果から，義務論一致による自動運転車の信頼への影響について，その効果はシナリオの内容によって変化する可能性について留意しておく必要があるだろう。

本研究は価値共有の効果の一般化可能性を示唆している。価値共有による信頼への効果はリスク認知研究において検討されてきたが，当然，信頼の対象となるのはリスク管理者であった。本研究では，人間が望む道徳判断と自動運転車や医療用人工知能が下す判断が一致することによって，そういった機械への信頼が高まることが示されていた。この知見は，価値の共有が，人間や機械に関わらず，信頼の規定因として機能することを示唆している。

しかし，現実場面において，道徳判断の一致によって，自動運転車への信頼を変化させることは難しいだろう。なぜなら，シナリオによって参加者の選好が変わるからである。例えば，Yokoi & Nakayachi (2021, *Hum Fac*) の実験 1 のシナリオでは約 8 割の参加者が功利主義を選好した。一方，実験 2 と 3 のシナリオでは，約 6 割から 8 割の参加者が義務論を選好した。このように，1 つのシナリオから人々の選好を理解することは難しい。自動運転車が下す道徳判断と使用者が望む道徳判断を一致させることも難しいだろう。

最後に本研究における限界点を述べる。それは，道徳判断の一致による信頼への効果がどの程度重要なのかという点である。これまでの自動運転車への信頼研究においては，機能や性能などの工学的な観点から信頼の要因が検討されてきた (e.g., Beller, Heesen, & Vollrath, 2013)。本研究では，機能や性能と比較して，道徳判断の一致がどれくらい信頼に影響を与えるのかを検討できていない。自動運転車が導入されたときに道徳判断の一致が信頼を決める要因としてどれくらい重要なのかを理解するためには，他の要因との比較が今後の課題となるだろう。