# “伦理旋钮”：破解无人驾驶算法困境的密钥？

张学义　王晓雪

［摘要］针对无人驾驶面临的算法困境，有学者提出了“伦理旋钮”算法理论，认为将无人驾驶汽车的算法设置权交付给车主，能够化解无人驾驶汽车在发生事故后所面临的道德—法律归责困境。但该算法可能会陷入集体性的“囚徒困境”：大多数车主为自保而在算法设置上作出极端利己的选择，致使社会总伤亡程度增加。为此，课题组运用实验哲学研究方法，将其与现有伦理算法进行了经验性比较，数据表明：“伦理旋钮”算法理论虽没有完全消解道德—法律归责困境，但其在无人驾驶汽车事故归责方面要比其他算法更加明晰，具有明显的稳定性；该算法理论上可能存在的“囚徒困境”在实际操作中并未出现，且具有更大的市场可行性。“伦理旋钮”算法理论为破解无人驾驶算法困境提供了可能方向，这是实验哲学研究方法在伦理学领域应用的有益尝试。

［关键词］伦理旋钮；无人驾驶；算法伦理；实验哲学

## 一、引言

随着大数据、人工智能等新兴技术的迅猛发展，自动驾驶将迎来其技术奇点，越来越多的资本与研发力量投入其中，并且在国内外开启了上路试点。有研究者不无乐观地指出，当度过 2020—2040 年的自动驾驶与传统人类驾驶共存的混合模式后，人们将在 2040 年后迎来全新的交通生态系统，并真正迈入无人驾驶新时代。①

然而，由于自动驾驶汽车将交通事故的核心问题由传统的责任承担与分配转向了自动驾驶汽车对生命安全的分配，其内生的伦理决策困境以及衍生出的事故责任归属模糊成为亟须解决的问题。在目前的自动驾驶情境中，这一困境尚可以被 3 级自动驾驶的人机共驾状态所消化。② 在该级别中，自动驾驶汽车的智能系统处理大部分驾驶工作，在遇到棘手的决策难题时将驾驶权交还给人类，这也是目前进行路测和运营的自动驾驶汽车配备一到两名人类安全员的主要理由。

作者：张学义，东南大学哲学与科学系副教授，江苏省道德发展智库研究员，zxynj0928@126.com；王晓雪，东南大学哲学与科学系 2017 级哲学专业本科生，yuki.xx9@qq.com。

＊　本文系国家社会科学基金重大项目“负责任的人工智能及其实践的哲学研究”（21&ZD063）、江苏省社会科学基金一般项目“当代科学理解的前沿问题研究”（22ZXB002）、东南大学研究生思政课程教改项目“《自然辩证法概论》课程分类教学模式改革创新研究”（yjgszkc2211）阶段性成果。北京大学哲学系博雅博士后研究人员隋婷婷在论文撰写过程中提供了诸多建设性意见和帮助，在此谨表诚挚谢意！

① 刘少山等：《第一本无人驾驶技术书》，Ⅶ页，北京电子工业出版社，2019。

② 按照我国的《汽车驾驶自动化分级》，自动驾驶汽车按自动化程度分为 6 个等级：0 级（应急辅助），1 级（部分驾驶辅助），2 级驾驶自动化（组合驾驶辅助），3 级（有条件自动驾驶），4 级（高度自动驾驶），5 级（完全自动驾驶）。

自动驾驶发展的终极方向是达到完全的无人驾驶，这一发展方向使得其在未来仍需脱离人车共驾模式。人车共驾也存在一定的安全隐患，当智能系统长久负责处理驾驶事宜，人类安全员将很难始终保持自身对于道路情况的注意力，若在遇到突发事件时瞬间被移交驾驶决策权，往往不太可能立刻收回注意力并作出理智的判断。即便安全员能够克服生理疲劳，保持注意力的高度集中，这似乎也违背了自动驾驶汽车将人类从传统驾驶活动中解放出来的初衷。因此，为自动驾驶汽车可能遇到的决策困境预设一种伦理性算法，并为随后的责任归属难题寻求一个普适性方案，便成为迫在眉睫的任务。

当前的算法按设计思路可分为两大类：强制型伦理算法（mandatory ethics setting，MES）和个人化伦理算法（personal ethics setting，PES）。前者是指自动驾驶车辆的设计师、制造商预先为车辆设置一个特定的道德算法，后者则是将算法的设置权交付给车主。[①] 面对伦理困境中涉及的功利主义与道义论的理论冲突，决策情境中并不存在一个超越所有原则的、最具正当性的伦理规范，不论选择何种伦理理论，都需面对其固有的缺陷。[②]

强制型算法以功利主义算法（utilitarian algorithm）、罗尔斯算法（Rawls algorithm）、制动力学算法（dynamics algorithm）等为主，但由于使用者对于伦理原则要求的非普适性，目前的强制型算法并不能满足人们实际操作的需要。个人化算法以伦理旋钮（ethical knob）为代表，该算法由贾斯帕·康提萨（Guiseppe Contissa）等人提出。伦理旋钮设置了一个由利他到利己的刻度旋钮（刻度为由 0 到 1），靠近 0 的一端对应了偏好乘客的利己主义模式，靠近 1 的一端对应了偏好行人的利他主义模式，旋钮中央则对应行人与乘客同等重要的公平模式，即在事故中保护人数较多的一方，若人数相等则随机选择。[③] 图 1 为伦理旋钮刻度表。

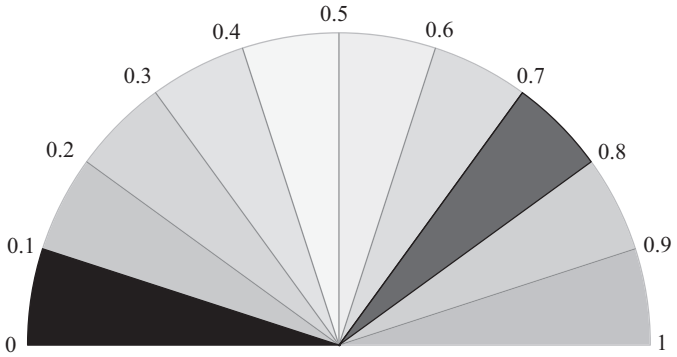


**图 1　伦理旋钮刻度表**

相比一元化的强制型算法，个人化算法为车主提供了相对多元的选择；同时，借由转让选择权使车主成为无人驾驶汽车的“道德代理”（moral proxy）[④]，似乎可以在一定程度上解决无人驾驶汽车决策主体模糊的问题，从而在事故归责方面比强制型算法更有优势。但简·果戈尔（Jan Gogoll）、朱利安·穆勒（Julian F. Müller）在对个人化算法的理论推演中指出，对于无人驾驶汽

① J. Gogoll，and J. F. Müller. “Autonomous Cars：In Favor of a Mandatory Ethics Setting”. *Science and Engineering Ethics*，2017，23（3）：681－700.

② 隋婷婷、张学义：《功利主义在无人驾驶设计中的道德算法困境》，载《自然辩证法研究》，2021（10）。

③ G. Contissa，et al. “The Ethical Knob：Ethically-customisable Automated Vehicles and the Law”. *Artificial Intelligence and Law*，2017，25（3）：365－378.

④ J. Millar. “Technology as Moral Proxy：Autonomy and Paternalism by Design”. *IEEE Technology and Society Magazine*，2015，34（2）：47－55.

车的道德决策所产生的伦理问题，更恰当的思考方式是博弈论，个人化伦理算法所带来的后果可能是车主在算法设置中陷入集体性的“囚徒困境”。[①]，换言之，在无人驾驶的个人化算法设置中，当人们无法确保其他的无人驾驶汽车车主会如何设定他们的伦理旋钮之时，出于自保，每个人都可能将自己的旋钮设置到极端利己主义状态，从而导致更多的伤害，进而丧失对整个无人驾驶汽车大环境的信任。[②]

本课题组采用实验哲学研究方法对伦理旋钮算法的理论推演进行了问卷调研，以验证伦理旋钮是否存在如理论假设的归责优势，即其是否能够有效消解道德—法律困境以及是否会出现车主博弈的囚徒困境。

## 二、伦理旋钮算法横向对比实验

本次实验分为三组（每组 230 名被试，共计 690 名，年龄涵盖 18～56 周岁人群），分别为罗尔斯算法 VS 伦理旋钮、制动力学算法 VS 伦理旋钮以及功利主义算法 VS 伦理旋钮。全部数据通过线上平台采集，样本数据来源覆盖全国大部分的省与直辖市。

问卷以自动驾驶的“行人难题”和“隧道难题”为基础情境进行实验，分别对比了伦理旋钮与强制化算法中较有代表性的罗尔斯算法、制动力学算法以及功利主义算法在归责、购买度等方面的异同。

“行人难题”来源于菲利帕·福特（Philippa Foot）提出的“电车难题”[③]，主要测试人们在不关涉自身安全性的情况下对行人的生命决策困境进行的道德选择：

> 你正在驾驶一辆汽车，道路前方出现了 5 个行人，道路侧前方有个可以转向的岔道，岔道上有 1 个行人。此时已经来不及刹车，直行可能会撞死 5 个行人，转向可能会撞死 1 个行人。这个时候，你会选择转向吗？

“隧道难题”则来源于简·果格尔提出的自动驾驶隧道两难情境，用于测试人们在关涉自身安全时的道德选择[④]：

> 你正在隧道里驾驶一辆汽车，道路前方出现了 5 个行人，没有可供你选择的其他路线。此时已经来不及刹车，汽车直行可能会撞死 5 个行人，如果不选择直行，只能转向撞向隧道墙壁，这样则可能会撞死你自己。这个时候，你会选择转向吗？

### （一）罗尔斯算法 VS 伦理旋钮算法

罗尔斯算法由德里克·里本（Derek Leben）根据罗尔斯《正义论》中的“最大化最小原则”（maximizing the minimum payoffs）和“无知之幕”（veil of ignorance）理论提出。“最大化最小原则”指决策者应考虑每种方案能造成的最糟糕的后果，在对比方案时，将所有方案的最坏结果进行排序，并以此为依据选择一个损失小于其他最坏结果的方案。[⑤]“无知之幕”则是使决策的人不知道自己所处的位置，以避免人们为自己所在的群体谋利。[⑥]因而，罗尔斯算法的关键点在于通过优先最大化底线安全的方式，首先筛选出事件相关人存活概率的最低收益集，再经由数据的循环穷举，筛选出使所有相关人存活率最大化的操作。这一算法只计算车祸相关人的存活

---

①②④ J. Gogoll, and J. F. Müller. “Autonomous Cars: In Favor of a Mandatory Ethics Setting”. *Science and Engineering Ethics*, 2017, 23 (3): 681－700, 681－700, 681－700.

③ P. Foot. “The Problem of Abortion and the Doctrine of Double Effect”. *Oxford Review*, 1967 (5): 5－15.

⑤⑥ J. Rawls . *A Theory of Justice*. Belknap Press of Harvard University Press, 1971, p. 154, p. 118.

概率，不涉及对车主或行人在身份和人数层面的偏袒，里本认为这是一种接近“无知之幕”的公平状态。①

在强制性算法的问卷情境里，罗尔斯算法在问卷中的行人难题情境被表述为：

如果你乘坐的无人驾驶汽车上安装了一套算法系统，该算法系统经过计算得出：汽车直行时，5人中受伤最严重的人死亡率为60%；汽车转向时，受伤最严重的人死亡率为90%。此时根据算法系统，该无人驾驶汽车选择了直行。

罗尔斯算法在问卷中的隧道难题情境被表述为：

该无人驾驶汽车上安装了一套算法系统，该算法系统经过计算得出：汽车直行时，5人中受伤最严重的人死亡率为60%；汽车转向撞墙时，你自己的死亡率为90%。此时根据算法系统，该无人驾驶汽车选择了直行，你是否认同这样的选择？

在伦理旋钮的问卷情境里，算法的伦理偏向由被试设置，在问卷中被表述为：

如果你乘坐的无人驾驶汽车安装了一个从0到1的连续刻度的算法旋钮，每个刻度都代表了无人驾驶汽车在遇到车祸时所作出的不同选择：旋钮转到中间，遇到危险时，将保护乘客与行人当中人数多的一方；旋钮越靠近0，越倾向于保护乘客，置于0时，该汽车将无视人数差异，绝对保护乘客；旋钮越靠近1，越是倾向于保护行人，置于1时，汽车将无视人数差异，绝对保护行人。无人驾驶汽车启动前，作为车主，你可以提前设定旋钮刻度，使得该汽车在遇到危险时作出自主的算法选择。

230人（女122，男108）参与了本次实验。不论是作为当事人还是作为旁观者，被试在被问及一旦出现车祸该如何归责时，伦理旋钮算法在行人难题与隧道难题情境中对于车主的归责均高于罗尔斯算法（不过作为旁观者，其归责车主的比例要略高于当事人视角）；在归责不清选项中，罗尔斯算法也普遍比伦理旋钮高（隧道难题的旁观者视角除外）。具体见表1。

**表1　　罗尔斯算法 VS 伦理旋钮归责情况**

| 归责主体 | 行人难题情境 | | | | 隧道难题情境 | | | |
|---|---|---|---|---|---|---|---|---|
| | 当事人视角 | | 旁观者视角 | | 当事人视角 | | 旁观者视角 | |
| | 罗尔斯算法 | 伦理旋钮算法 | 罗尔斯算法 | 伦理旋钮算法 | 罗尔斯算法 | 伦理旋钮算法 | 罗尔斯算法 | 伦理旋钮算法 |
| 车主 | 32.61% | 45.65% | 41.74% | 52.17% | 37.83% | 50.43% | 46.52% | 50.43% |
| 生产商 | 22.17% | 20.43% | 25.22% | 20.87% | 23.04% | 19.13% | 20.00% | 16.96% |
| 设计人员 | 22.61% | 21.30% | 15.65% | 13.04% | 22.61% | 15.22% | 18.26% | 15.65% |
| 不清楚 | 22.61% | 12.61% | 17.39% | 13.91% | 16.52% | 15.22% | 15.22% | 16.96% |
| 卡方检验 | $X^2=131.223$，df=9，$P<0.01$ | | $X^2=189.885$，df=9，$P<0.01$ | | $X^2=131.223$，df=9，$P<0.01$ | | $X^2=176.814$，df=9，$P<0.01$ | |

在伦理旋钮的囚徒困境测试中，在行人难题与隧道难题情境中，不管是当事人视角还是旁观者视角，被试均未过度选择极端利己主义的“0”刻度；相反，更多选择了偏向中立的“0.5”刻度。这一结果预示，在实际的操作中，大多数人在乘坐装置伦理旋钮的无人驾驶汽车遇到危险时，并不会作出极端利己的选择，而是采取相对公平的选项，果戈尔等人在理论上推演的囚徒困境现象在实测中并未出现。具体见表2。

① D. Leben. “A Rawlsian Algorithm for Autonomous Vehicles”. *Ethics and Information Technology*, 2017, 19 (2): 107－115.

**表 2**　　**罗尔斯算法 VS 伦理旋钮算法之囚徒困境**

| 囚徒困境 | 行人难题 | | 隧道难题 | |
|---|---|---|---|---|
| 旋钮刻度 | 当事人视角 | 旁观者视角 | 当事人视角 | 旁观者视角 |
| 0 | 8.7% | 4.3% | 8.7% | 2.2% |
| 0.1 | 8.3% | 1.3% | 7.4% | 3.0% |
| 0.2 | 6.1% | 3.0% | 7.0% | 4.3% |
| 0.3 | 9.6% | 2.6% | 12.2% | 1.7% |
| 0.4 | 10.4% | 5.2% | 9.1% | 4.3% |
| 0.5 | 27.8% | 29.6% | 27.4% | 34.3% |
| 0.6 | 5.7% | 7.4% | 7.8% | 7.0% |
| 0.7 | 4.8% | 6.1% | 4.3% | 9.1% |
| 0.8 | 3.9% | 10.4% | 4.3% | 7.0% |
| 0.9 | 5.2% | 11.3% | 3.9% | 8.3% |
| 1 | 9.6% | 18.7% | 7.8% | 18.7% |

此外，在两个情境的购买欲求对比中，人们对装置伦理旋钮算法的无人驾驶汽车的购买意向也略高于装置了罗尔斯算法的无人驾驶汽车，且在统计学意义上呈现出差异性显著；不过，被试对装置两种算法的无人驾驶汽车的真正购买欲求普遍都不高。具体见表 3。

**表 3**　　**罗尔斯算法 VS 伦理旋钮算法之购买欲求**

| 购买欲求 | | 不会 | 会 | 卡方检验 |
|---|---|---|---|---|
| 行人难题 | 罗尔斯算法 | 57.4% | 42.6% | $X^2=73.849$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 51.7% | 48.3% | |
| 隧道难题 | 罗尔斯算法 | 54.8% | 45.2% | $X^2=114.649$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 50.9% | 49.1% | |

### （二）制动力学算法 VS 伦理旋钮算法

制动力学算法是基于减少车辆在失控状态下造成侧滑、翻滚等风险而提出的技术型算法。由于汽车转向时轮胎的角度和速度的变化是不可避免的，但失控时轮胎角度和速度的变化极可能使轮胎与地面的动摩擦变为静摩擦，使得汽车侧翻或旋转，并引起爆炸等破坏性后果。同时，考虑到现实生活中，行人可能会作出躲避等动作，汽车急转弯撞人的概率会高于车辆保持原有行车轨迹时的概率。① 因而，这一算法认为无人驾驶车辆在紧急状态下降低破坏力的优选方式是选择直行，可以减少车辆失控带来的连带伤害。

在强制性算法的问卷情境里，制动力学算法被表述为：

> 如果你乘坐的无人驾驶汽车上安装了一套算法系统，该算法系统经过计算得出：汽车转向撞击行人可能造成的伤害性更大，甚至可能造成翻车或其他不可控的连带事故，而选择直行可能造成的伤害性最小。该无人驾驶汽车选择了直行。

230 人（女 112，男 118）参加了本次实验。在事故归责方面，被试不论是作为当事人还是作为旁观者，伦理旋钮算法在行人难题与隧道难题情境中对于车主的归责选择均高于制动力学算法（只是作为旁观者，其归责车主的比例仍略高于当事人视角）；在归责不清选项中，制动力学算法也普遍比伦理旋钮高（隧道难题的旁观者视角除外）。具体见表 4。

① R. Davnall. “Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics”. *Science and Engineering Ethics*，2019，4（1）：1353－3452.

**表 4　　制动力学算法 VS 伦理旋钮算法归责情况**

| 归责主体 | 行人难题情境 | | | | 隧道难题情境 | | | |
|---|---|---|---|---|---|---|---|---|
| | 当事人归责 | | 旁观者归责 | | 当事人归责 | | 旁观者归责 | |
| | 制动力学算法 | 伦理旋钮算法 | 制动力学算法 | 伦理旋钮算法 | 制动力学算法 | 伦理旋钮算法 | 制动力学算法 | 伦理旋钮算法 |
| 车主 | 32.61% | 47.83% | 42.17% | 56.52% | 34.35% | 42.61% | 46.52% | 54.78% |
| 生产商 | 23.04% | 19.13% | 21.74% | 18.26% | 26.52% | 23.48% | 23.48% | 18.70% |
| 设计人员 | 27.83% | 21.74% | 22.17% | 15.22% | 27.39% | 23.48% | 19.57% | 15.22% |
| 不清楚 | 16.52% | 11.30% | 13.91% | 10.00% | 11.74% | 10.43% | 10.43% | 11.30% |
| 卡方检验 | $X^2=125.801$，df=9，$P<0.01$ | | $X^2=133.338$，df=9，$P<0.01$ | | $X^2=195.408$，df=9，$P<0.01$ | | $X^2=228.183$，df=9，$P<0.01$ | |

在伦理旋钮的囚徒困境测试中，被试在行人难题与隧道难题中，不论是当事人视角还是旁观者视角，大多数人并未选择完全利己的“0”刻度。具体见表 5。

**表 5　　制动力学算法 VS 伦理旋钮算法之囚徒困境**

| 囚徒困境 | 行人难题 | | 隧道难题 | |
|---|---|---|---|---|
| 旋钮刻度 | 当事人视角 | 旁观者视角 | 当事人视角 | 旁观者视角 |
| 0 | 8.3% | 4.8% | 9.1% | 3.9% |
| 0.1 | 6.1% | 2.2% | 6.1% | 3.9% |
| 0.2 | 6.5% | 2.2% | 6.1% | 1.3% |
| 0.3 | 6.5% | 3.9% | 7.8% | 3.0% |
| 0.4 | 9.1% | 3.9% | 8.7% | 3.5% |
| 0.5 | 28.7% | 30.0% | 29.6% | 30.9% |
| 0.6 | 7.4% | 10.0% | 6.1% | 8.3% |
| 0.7 | 6.5% | 4.8% | 7.0% | 7.4% |
| 0.8 | 4.8% | 7.4% | 5.7% | 8.7% |
| 0.9 | 3.9% | 10.0% | 3.9% | 10.9% |
| 1 | 12.2% | 20.9% | 10.0% | 18.3% |

在上述两个情境的购买欲求对比中，人们对安装了伦理旋钮的无人驾驶汽车的购买意向依然略高于安装了制动力学算法的汽车，并在统计学层面上有显著性差异；同样，被试对装置此两种算法的无人驾驶汽车的真正购买欲求普遍都不高。具体见表 6。

**表 6　　制动力学算法 VS 伦理旋钮算法之购买欲求**

| 购买欲求 | | 不会 | 会 | 卡方检验 |
|---|---|---|---|---|
| 行人难题 | 制动力学算法 | 54.3% | 45.7% | $X^2=68.202$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 48.3% | 51.7% | |
| 隧道难题 | 制动力学算法 | 58.3% | 41.7% | $X^2=85.967$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 49.1% | 50.9% | |

### （三）功利主义算法 VS 伦理旋钮算法

功利主义算法作为传统电车难题在大众直觉中最具普适性的答案，是道德算法早期最受期待的算法之一。该算法通常以人数为决策的基础，在事故中保护多数人，牺牲少数人。

在强制性算法的问卷情境里，功利主义算法被表述为：

> 如果该无人驾驶汽车上安装了一套算法系统，该算法系统的计算原则是：汽车在遇到危险时，总是以牺牲少数人的生命以拯救多数人的生命的方式来行驶；而此时选择转向可能撞死那 1 个行人，选择直行则可能会牺牲那 5 个行人。该无人驾驶汽车选择了转向。

230 人（女 116，男 114）参加了本次实验。无论是当事人视角还是旁观者视角，被试在行人难题与隧道难题情境中对于车主归责的选择时，伦理旋钮算法的选项都高于功利主义算法的选项，并且依然呈现出旁观者视角高于当事人视角的状况。具体见表 7。

**表 7　　功利主义算法 VS 伦理旋钮算法归责情况**

| 归责主体 | 行人难题情境 | | | | 隧道难题情境 | | | |
|---|---|---|---|---|---|---|---|---|
| | 当事人归责 | | 旁观者归责 | | 当事人归责 | | 旁观者归责 | |
| | 功利主义算法 | 伦理旋钮算法 | 功利主义算法 | 伦理旋钮算法 | 功利主义算法 | 伦理旋钮算法 | 功利主义算法 | 伦理旋钮算法 |
| 车主 | 30.43% | 42.61% | 40.43% | 48.26% | 30.00% | 40.87% | 36.09% | 46.09% |
| 生产商 | 23.48% | 24.35% | 23.04% | 22.17% | 30.00% | 23.91% | 28.26% | 23.48% |
| 设计人员 | 16.52% | 16.52% | 13.91% | 14.35% | 20.43% | 16.96% | 17.83% | 15.22% |
| 不清楚 | 29.57% | 16.52% | 22.61% | 15.22% | 19.57% | 18.26% | 17.83% | 15.22% |
| 卡方检验 | $X^2=123.667$，df=9，$P<0.01$ | | $X^2=142.198$，df=9，$P<0.01$ | | $X^2=262.345$，df=9，$P<0.01$ | | $X^2=188.608$，df=9，$P<0.01$ | |

在伦理旋钮的囚徒困境测试中，不论是当事人视角还是旁观者视角，被试在行人难题与隧道难题情境中，依然未出现大多数人选择完全利己刻度“0”的现象。具体见表 8。

**表 8　　功利主义算法 VS 伦理旋钮算法之囚徒困境**

| 囚徒困境 | 行人难题 | | 隧道难题 | |
|---|---|---|---|---|
| 旋钮刻度 | 当事人视角 | 旁观者视角 | 当事人视角 | 旁观者视角 |
| 0 | 12.6% | 3.9% | 11.3% | 3.5% |
| 0.1 | 4.8% | 2.2% | 5.7% | 2.2% |
| 0.2 | 7.0% | 1.7% | 7.8% | 3.9% |
| 0.3 | 9.6% | 2.6% | 10.9% | 3.5% |
| 0.4 | 6.5% | 3.5% | 7.0% | 3.0% |
| 0.5 | 30.4% | 32.2% | 29.6% | 30.9% |
| 0.6 | 5.7% | 5.2% | 4.8% | 7.0% |
| 0.7 | 5.2% | 6.1% | 5.2% | 7.0% |
| 0.8 | 4.3% | 10.4% | 4.3% | 9.1% |
| 0.9 | 5.7% | 8.3% | 3.5% | 7.8% |
| 1 | 8.3% | 23.9% | 10.0% | 22.2% |

在购买欲求对比中，人们对设置了伦理旋钮无人驾驶汽车的购买意向依然略高于设置了功利主义算法的无人驾驶汽车，并在统计学层面有显著性差异。具体见表 9。

**表 9　　功利主义算法 VS 伦理旋钮算法之购买欲求**

| 购买欲求 | | 不会 | 会 | 卡方检验 |
|---|---|---|---|---|
| 行人难题 | 功利主义算法 | 48.7% | 51.3% | $X^2=117.274$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 45.2% | 54.8% | |
| 隧道难题 | 功利主义算法 | 53.0% | 47.0% | $X^2=136.999$，df=1，$P<0.01$ |
| | 伦理旋钮算法 | 47.0% | 53.0% | |

## 三、伦理旋钮的算法意义

### （一）伦理旋钮的归责优势

无人驾驶汽车一直存在道德—法律责任难以界定的问题。通常认为，在现有技术条件下，作为

机器，无人驾驶汽车并不具备完全意义上的主体性，即：不论是传统主体性观点谈及的能动性、具身意义的封闭性和意向性特征，还是非标准观点所说的概念化能力、因果推理能力、反事实推理能力和语义能力等条件，作为一种技术智能体的无人驾驶汽车，在当下乃至未来相当长的时间内都无法完全达到其要求。退一步来讲，即使无人驾驶汽车具备了主体性资格，那么，它是否拥有道德—法律能力呢？所谓道德—法律能力，是指无人驾驶汽车在发生交通事故时，它知道在道德、法律上应该做何选择，并根据道德原则、法律规范作出决策、采取行动，同时还能够对自己的行为进行合理的解释等。① 这些道德—法律能力对于现有的抑或是按照现有的技术进路进行研发的无人驾驶汽车而言是无法具备的。基于此，无人驾驶汽车既不具有完全意义上的道德—法律主体性资格，也无法拥有充分的道德—法律能力，更不能承担相应的道德—法律责任，“出于社会心理，人们不能接受一旦发生事故却没有特定的责任人，而可能是一台机器为此承担责任”②。

伴随着算法技术不断向更加智能状态进化发展，智能机器经历了由人类的行为投射发展到依靠算法进行自我深度学习的过程。由于算法决策的结果涉及人的生命安全在交通事故中的分配，因此，无论是法律责任还是道德责任，都必须被归结到某个具体的决策主体上，也即人们必须为无人驾驶汽车的决策结果寻找一个能够担保并承担责任的人。而将这一责任简单地归咎于厂商或是算法设计师显然都不太现实，这也将极大地打击或削弱无人驾驶研发者的创新动力。伦理旋钮算法为这个问题的解决提供了一个方向——将使用者拉入决策的过程。伦理旋钮的算法可选性一方面保留了个体在伦理困境决策中道德偏好的多样性，尊重了生命伦理当中不将个人道德选项强加于他人的共识③；另一方面，也通过让车主对算法的重置完成了对决策权的转渡。

虽然在问卷结果中，伦理旋钮算法对于车主的归责率在50%左右，并未完全化解无人驾驶汽车所面临的道德—法律困境，但其稳定性均高于其他算法；在不同的对照组实验中，在前置算法情境不同的情况下仍然保持了较高的稳定性和容错性。这在一定程度上代表了大众直觉对伦理旋钮的判断有着恒常性。

此外，伦理旋钮算法所作出的决策，不单纯是程序工程师对于无人驾驶汽车的算法预置，而是基于使用者自身的责任选择所最终达成的个性化算法。这在一定程度上避免了寻责无果的情况，对决策型算法的后续改进与发展提供了一个很好的参考。

### （二）伦理旋钮的囚徒困境

针对上述可能出现的困境，本次调查的数据反映，这一理论上的推演在实测层面并没有想象得那么严重。在大众选择中，尽管有10%左右的被试选择了极端利己设置（行人在算法中没有任何权重，汽车将在车祸困境中无条件保护车主，即便此时转弯避让行人可能只导致车主轻伤），但选项中占比最大的（30%左右）仍是中立（0.5）的算法设置。也就是说，虽然存在极端利己的选择，但实质上有别于果戈尔认为的大多数人都将遵循囚徒困境作出极端利己设置的推测。90%的人在问卷中均选择给予行人不同的算法权重（即根据权重计算选择是避让还是撞向行人），能够在很大程度上保证行人的生还率。同时，考虑到人类驾驶情境中同样会有一部分人在车祸中作出极端利己的选择，算法设置上10%的极端利己似乎仍在可接受的范围。由此可见，算法对人们选择的影响与操纵并没有达到一个不可控制的状态，起码对于安装了伦理旋钮可能引发的囚徒困境来说便是如此。

---

① 吴童立：《人工智能有资格成为道德主体吗?》，载《哲学动态》，2021（6）。

② 白惠仁：《自动驾驶汽车的“道德算法”困境》，载《科学学研究》，2019（1）。

③ J. Millar. “Technology as Moral Proxy: Autonomy and Paternalism by Design”. *IEEE Technology and Society Magazine*, 2015, 34（2）: 47－55.

此外，虽然算法决策是以数据分析与实施的判断为基础的，但决策的过程仍需要加入伦理、道德、情感等方面的因素，分析算法本身是否有缺陷与偏见，以便有效运用算法系统对人们有所助益的方面。在可以预见的未来，人工智能伦理算法也许能够帮助人们更好地摆脱伦理困境，但完全将伦理决策交给机器，尚无法得到完全的信赖。一方面，放任智能系统全权代劳人类主体的决策活动，挤压了个体对于价值性问题的自主决策空间，可能造成人类的主体性和自治性在智能时代被自动化决策所消解；另一方面，由于个体在价值观、道德原则等方面的差异，普适性的强制型算法原则可能永远也无法找到，因而在道德允许的范围内给予无人驾驶汽车用户以算法选择权，可能是目前最具可行性、也能最大程度保留使用者主体性的方案。

**（三）伦理旋钮的市场接受度**

伦理旋钮算法理论的提出者认为，伦理旋钮由于其能够由乘客设定的特点而会具有更高的市场接受度。在本次实验中，大众对伦理旋钮的购买意向略高于 50%。尽管优势不够明显，但相比购买意向在 40%左右的罗尔斯、制动力学以及功利主义等算法，其表现仍算是可圈可点。在没有其他更好的算法产生之前，伦理旋钮不失为一个优于现有其他算法的选项。

如果要求所有的无人驾驶汽车都强制采用并严格执行公正的（功利主义）伦理算法，即：遭遇危险时，总是以牺牲少数人来拯救多数人，许多人仍然可能拒绝购买、乘坐无人驾驶汽车，即使无人驾驶相对于人工驾驶在安全性方面具有更加明显的优势。此外，如果无人驾驶汽车固定的伦理算法设置交由生产者抉择，那么生产者可能会更多地考虑车主或乘客是否会购买或租用，从而采用更加偏向保护乘客的伦理算法。然而，如果伦理算法倾向于保护乘客的安全，那么，无人驾驶汽车遇到危险时，将不可避免地首选对行人造成伤害，致使行人处于危险之中；而安装了如此算法的无人驾驶汽车恐怕也无法获得相关部门的上路许可。

而搭载伦理旋钮算法的无人驾驶汽车，车主或乘客有责任决定在突发事故情况下应采用何种伦理原则，并且能够对他们作出的选择负责；同时，无人驾驶汽车也能够基于风险评估来执行用户的道德选择，这对于提高人们对自动驾驶汽车的接受度、加速自动驾驶汽车的市场投放等都是有利的。

## 四、结语

**（一）伦理旋钮算法是否能破解无人驾驶算法困境？**

通过对伦理旋钮算法的综合分析以及与其他伦理算法在解决无人驾驶伦理困境中的对比表现来看，伦理旋钮算法均有着较好的应用效果：通过平行对比现有算法模式的实际效果，验证了伦理旋钮算法在道德—法律归责、购买欲求以及市场接受度等方面具有较为明显的优势；同时，数据也表明，基于理论推演可能存在的囚徒困境缺陷在实验中也并未出现。

在购买方面，人们对伦理旋钮算法的接受度和对搭载了伦理旋钮算法的无人驾驶汽车的购买欲求均高于其他算法。在归责方面，虽然伦理旋钮并没有完全消解道德—法律困境，但在无人驾驶汽车事故归责方面相比其他伦理算法更加明晰，且能保持在一个相对稳定的状态。伦理旋钮算法将用户拉入责任选择的运算过程当中，在一定程度上避免了责任主体缺失的情况。此外，伦理旋钮的可调节性，也使得无人驾驶汽车在进入市场之后能够根据乘坐成员的不同，设置更加多元化、更加适配的算法选项，为适应不同主体的不同需求提供了可能性。尽管我们可以假设未来的人工智能伦理算法也许能够更好地解决伦理困境，但人们的个体化选择权依然重要，算法的自主可选性是发挥使用者主体性、避免其被科技消解的方式之一。在缺陷方面，伦理旋钮算法理论上可能存在类似囚徒困境的大规模恶性博弈场景在实验中并不明显，使用者们大多都选择了比较中立的算法调节方案。诚然，伦理旋钮极端利己主义模式的存在也许会导致总的死亡率比强制性统一预设的伦理算法更

高，但若只有少部分群体进行了这一设置，对于整体的驾驶群体似乎尚可接受。在现实中，可以通过法律、制度等途径对实施极端利己主义模式的行为作出严格规定和限制，辅以高额的保险代价和更小的保险范围，理论推演中的极端利己主义选项比例可能会大幅减少。

因此，我们认为，个人化定制伦理算法在无人驾驶汽车伦理算法的后续发展中卓有优势，与之相关的法律规定和社会规则也可为无人驾驶汽车投入和适应社会作出补充规定来约束个人化定制算法设定，从而避免其缺陷所可能造成的社会不良影响。

**（二）实验哲学研究方法的意义**

不难看出，伦理旋钮的实验哲学研究为无人驾驶伦理算法如何落地提供了可能的启示性方案，也为传统伦理学的理论研究如何借助实验哲学方法进行哲学探讨提供了鲜活的案例。以直觉作为切入口的实验哲学，通过将抽象的哲学理论或观点具化为描述性经验情境，并诉诸大众直觉而非哲学家私人化的神秘内省，从而对常常是基于哲学家个人经验而提炼出来的哲学理论或观点进行验证，提供或支持或反驳的经验性论据，进而对该议题提出更加合理、规范的要求，抑或是通过经验性的证据启迪理论探讨的新维度。①

总之，实验哲学并非要取代、也不可能取代纯粹思辨的扶手椅哲学，而是为其配备了一套助探式的"新工具"，为哲学家进行哲学探索提供了一种可供选择的新途径，开启了一条"坐而论道"与"起而践道"协同共进的新路径。

## Ethical Knob: The Key to Solve the Dilemma of Driverless Algorithms?

ZHANG Xueyi, WANG Xiaoxue

(Department of Philosophy and Science, Southeast University)

**Abstract**: In viewing of the algorithm dilemma faced by driverless cars, Italian scholar Giuseppe Contissa and his colleagues proposed the theory of "ethical knob" algorithm: the right to set the algorithm of driverless cars should be handed over to the owner, so as to solve the moral-legal liability dilemma caused by driverless cars when they cause accidents. However, this algorithm may fall into the collective "prisoner's dilemma:" most car owners make extremely selfish choices in the algorithm setting for self-protection, resulting in an increase in the total social casualties. This article uses the methods of experimental philosophy to compare "ethical knob" with current ethical algorithms, to verify people's acceptance of various ethical algorithms and their desire to purchase self-driving cars equipped with such ethical algorithms. The data show that although the ethical knob does not completely solve the moral-legal liability dilemma, it is clearer and stabler than other algorithms in terms of the liability of driverless vehicle accidents. The "prisoner's dilemma" that may exist in this algorithm does not appear in practice. Moreover, driverless cars equipped with ethical knobs have greater market viability. The theory of "ethical knob" algorithm provides a possible solution to solve the dilemma of driving algorithm, and it is also a beneficial attempt to apply experimental philosophy research methods in the field of ethics.

**Key words**: Ethical knob; Driverless; Algorithm ethics; Experimental philosophy

（责任编辑　李　理）

① 聂敏里：《哲学与实验——实验哲学的兴起及其哲学意义》，载《自然辩证法通讯》，2020（9）。