自动化信任的研究综述与展望

董文莉 [1] 方卫宁 [1]

摘 要 随着自动化能力的快速提升, 人机关系发生深刻变化, 人的角色逐渐从自动化的主要控制者转变为与其共享控制的合作者. 为了实现绩效和安全目标, 人机协同控制需要操作人员适当地校准他们对自动化机器的信任, 自动化信任问题已经成为实现安全有效的人机协同控制所面临的最大挑战之一. 本文回顾了自动化信任相关文献, 围绕自动化信任概念、模型、影响因素及测量方法, 对迄今为止该领域的主要理论和实证工作进行了详细总结. 最后, 本文在研究综述和相关文献分析的基础上提出了现有自动化信任研究工作中存在的局限性, 并从人机系统设计的角度为未来的自动化信任研究提供一些建议.

关键词 自动化信任, 信任校准, 人机协同控制, 人机系统设计

引用格式 董文莉, 方卫宁. 自动化信任的研究综述与展望. 自动化学报, 2021, **47**(6): 1183−1200

DOI 10.16383/j.aas.c200432

Trust in Automation: Research Review and Future Perspectives

DONG Wen-Li[1] FANG Wei-Ning[1]

Abstract With the rapid improvement of automation capability, the human-machine relationship has undergone profound changes. The role of human has gradually changed from the main controller of automation to the partner sharing control with it. Human-machine collaborative control requires the human operator to appropriately calibrate their trust in automatic machine in order to achieve performance and safety goals. Trust in automation has proved to be one of the greatest challenges to achieve safe and effective human-machine collaborative control. This paper reviews the literature related to trust in automation, and summarizes the main theoretical and empirical work in this field up to now in detail, centering on the concepts, models, influencing factors and measurement methods of trust in automation. Finally, this paper explains limitations that are present in existing research works based on research review and relevant literature analysis, and provides some suggestions for future research on trust in automation from the point of human-machine system design.

Key words Trust in automation, trust calibration, human-machine collaborative control, human-machine system design

Citation Dong Wen-Li, Fang Wei-Ning. Trust in automation: Research review and future perspectives. *Acta Automatica Sinica*, 2021, **47**(6): 1183−1200

得益于人工智能 (Artificial intelligence, AI) 技术的发展, 自动化的能力大幅提升. 许多自动化机器都具有一定的自主能力, 它们能够执行诸如计划和决策等复杂的高级认知任务, 它们有能力成为人类的合作伙伴, 甚至取代人类. 自动化的应用也达到了前所未有的广度和深度. 自动化机器被广泛部署到各种工作环境之中, 自动驾驶汽车、自主机器人以及决策辅助设备等已经被广泛集成到军事[1]、交通运输[2]、过程控制[3] 及医疗保健[4] 等领域; 自动化机器被更加多样化的最终用户群体所使用, 最具代表性的就是类似于自动驾驶汽车的驾驶员这类快速增长的非专业用户[5]. 在其工作环境中, 人与自动化机器协同控制的效能得到不断改善, 人机协作系统变得更加安全有效.

尽管如此, 由于作为智能自动化系统基础的 AI 技术[6] 的不足, 自动化的发展及应用面临着一些严峻的问题. 从 AI 的发展历史来看, 第二次 AI 技术浪潮目前正处于全盛阶段, 它由执行统计对象识别以及在大量数据中寻找模式的统计系统组成, 典型范例如人工神经网络系统, 这些统计系统已经在多个领域取得重大进展[7]. 然而, 以 “统计学习” 为特征的 AI 系统存在固有缺陷. 首先, 虽然这些系统对特定问题具有较强的推理和判断能力, 但它们没有实时的交互式学习能力, 不能处理动态目标和情

收稿日期 2020-06-17 录用日期 2020-09-07
Manuscript received June 17, 2020; accepted September 7, 2020
北京市自然科学基金 (L191018) 资助
Supported by Beijing Natural Science Foundation (L191018)
本文责任编委 曹向辉
Recommended by Associate Editor CAO Xiang-Hui
1. 北京交通大学轨道交通控制与安全国家重点实验室 北京 100044
1. State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044

境[8]. 因此, 在不久的将来实现在动态和非结构化环境中运行的完全自主的自动化系统是非常困难的, 并且, 出于道德考虑和责任需要, 人们可能不希望赋予自动化系统完全的自主权[9]. 其次, 作为统计系统核心的机器学习模型和学习过程不透明且输出结果难以解释[10], 这使得用户尤其是非专业用户理解自动化变得非常困难.

完全自主的自动化系统在未来很长一段时间内都不会出现, 自动化系统的主要作用仍然是作为人机团队成员与人共享控制而不是取代人, 人将继续参与或至少在某种程度上参与自动化系统的决策循环. 目前的人机协同控制主要遵循两种交互范式: 决策支持和监督控制[11]. 决策支持是指自动化为操作者提供可能的选择, 而监督控制则是指操作者监督自动化运行并在其失效或出现意外事件时及时接管控制权进行适当干预. 在这两种交互范式中, 操作者在很大程度上是基于他们对自动化意图及行为的理解做出正确选择. 然而, 由于自动化系统所采取的底层技术的限制, 期望操作者完全理解其自动化伙伴是不切实际的. 人难以理解自动化机器已经成为限制人机协同控制安全实施及效能发挥的一个瓶颈, 如何解决该问题从而将人的认知能力与自动化机器的计算能力紧密结合来实现更加安全有效的人机协同控制成为目前自动化机器开发和部署的难点.

既然人对自动化的理解与自动化的实际能力之间总是存在差距, 人缺乏客观评估自动化能力的缺陷只能用信任来弥补[12]. 信任是发展有效关系的关键因素, 信任在人类合作中的重要性也得到了广泛认可[13], 自动化信任 (Trust in automation), 即人对自动化的信任, 已经被确定为调节人与自动化之间关系的关键因素[14], 其作用方式与人类之间的信任相似[15]. 操作者的自动化信任水平与自动化的实际能力之间的匹配关系称为自动化信任校准[16], 如图 1 所示. 自动化系统本身的复杂性使得操作者几乎总是处于不当的自动化信任校准状态, 要么高估自动化的能力对其过度信任, 不加判别地依赖自动化导致误用; 要么低估自动化的能力对其缺乏信任, 导致停用[17]. 误用和停用都可能导致灾难性的后果. 例如, 2018 年 3 月, 一辆特斯拉汽车的车主去世, 事故调查将该起事故归咎于驾驶员对自动驾驶汽车的过度信任[18]. 特斯拉透露, 在自动驾驶仪没有识别出道路前方的混凝土障碍物并加速撞上障碍物之前, 驾驶员有足够的时间进行干预, 以防止撞车, 但他没有采取行动. 在随后对该起事故的补充说明中, 特斯拉认为, 与非自动驾驶汽车相比, 自动驾驶汽车的安全性预计将提升 10 倍, 如果公众对自动驾驶汽车缺乏信任并因此拒绝使用, 自动驾驶汽车可靠性提高所带来的安全性提升将无法实现, 这将造成每年全世界约有 90 万本可以被挽救的生命因此损失[19]. 操作人员不能正确使用自动化对人机协同控制的有效性和安全性造成巨大损害, 因此, 自动化信任问题应该至少与技术问题受到同等程度的重视. 在自动化设计和部署过程之中着重考虑自动化信任问题, 设计具有信任意识的人机系统至关重要, 它有助于人机系统最大限度地发挥人与自动化机器的潜力, 改善人机协同控制的绩效和安全性.

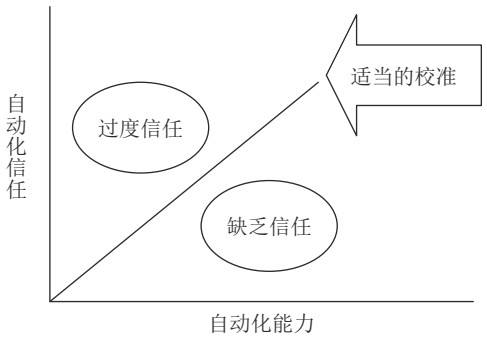


图 1 自动化信任校准示意图

Fig. 1 Diagram of calibration of trust in automation

1 自动化信任的研究现状

目前, 自动化信任相关研究主要集中在四个方面: 自动化信任概念与内涵、自动化信任模型、自动化信任影响因素、自动化信任测量方法. 自动化信任研究的目标是在明确自动化信任概念及内涵的基础之上, 通过构建自动化信任模型、探究自动化信任影响因素、发展自动化信任测量方法, 最终实现合适的自动化信任校准. 因此, 本文按照上述四个方面对迄今为止自动化信任研究的主要理论及实证工作进行综述.

1.1 自动化信任概念与内涵

明确相关概念与内涵是开展自动化信任理论及实证研究工作的基础. 已有文献对自动化概念与内涵的探究经历了以下三个阶段: 1) 总结其他领域的信任研究, 例如, 人际信任[20−21]、虚拟团队中的信任[22]、电子商务中的信任[23−25]、对信息系统的信任[26] 等, 综合人机交互背景下的相关理论以及实证研究成果, 探索自动化信任的本质, 明确自动化信任定义及相关概念, 发展自动化信任概念体系[14, 27]; 2) 总结人机交互领域的信任研究, 例如, 人−计算机交互中的信任[28−29]、人−机器人交互中的信任[30−31]、对决策支持系统的信任[32]、对车辆技术的信任[33] 以及对人工

智能算法的信任[34] 等, 在自动化信任概念及实证研究成果的基础上, 对已有自动化信任概念体系进行补充和完善[35−36]; 3) 在一般自动化信任概念体系的基础上, 综合特定人−自动化交互的特点, 如人−机器人交互、人−自动驾驶汽车交互等, 发展特定类型的自动化信任概念体系[30]. 三种类型的自动化信任概念研究相辅相成, 层层递进, 向发展明确的、可操作的自动化信任概念体系的目标逐步前进, 并为其他方面的自动化信任研究奠定理论基础. 下面, 本文将对一般自动化信任概念及内涵进行阐述.

1.1.1　自动化

自动化的发展和复杂化使自动化信任对人机协同控制效能的影响日益突出, 明确自动化信任概念与内涵需要建立在充分理解自动化的基础之上.

自动化信任文献中已经出现了常用的自动化定义. Lee 和 See 总结了自动化执行任务的过程, 将自动化定义为 “主动选择数据、转换信息、做出决策或控制流程的技术”[27], 该定义在自动化信任研究中得到广泛应用.

然而, 当前自动化信任文献中的自动化内涵并不清晰. 在自动化信任主题相关文献中, 术语自动化 (Automation)、自治 (Autonomy) 和自主系统 (Autonomous system) 经常被互换使用, 但它们的含义是存在差异的, 明确三者之间的关系是深入理解自动化内涵的关键.

自动化和自治的目的都是在很少或没有人工干预的情况下完成任务, 但早期自动化系统通常采用基于逻辑的编程, 在实现这一目标的能力上是有限制的; 自主系统则是基于计算智能和学习算法, 根据操作和情境信息的输入进行学习和进化以更好地适应不断变化的情况, 可以在不需人工干预的情况下长期、良好地完成目标[11], 但是随着时间的推移, 系统的行为也必然变得更加不确定. 自主系统自治水平的提高依赖于其自动化等级的增加, 一个特定的自主系统可以包含部分或全部的自动化等级[37]. 然而, 实现完全的系统自治是相当困难的, 在未来很长一段时间内, 大多数自主系统都将以某种程度的半自治 (Semi-autonomy) 存在, 即只有系统的某些方面发展为自治. 因此, 自动化和自治这两个术语并不存在一定形式的对立或矛盾, 它们代表了连续的技术进化的不同阶段[38], 自治是自动化的后期发展, 自主系统是能力更强且更加复杂的自动化系统. 本文综述的自动化信任研究所涉及的自动化对象是指处于半自治程度的自动化系统.

1.1.2　自动化信任

信任在众多领域广泛存在. 信任及其在合作关系中的调节作用已经在心理学、社会学、哲学、政治学、经济学等领域得到广泛研究, 在不同的研究领域存在超过 300 个信任定义[39]. 然而, 信任是一个复杂而模糊的概念, 目前研究者们对信任的定义尚无共识[40]. 如何定义信任对于研究自动化信任具有重要的理论和实践意义, 不一致的信任描述会导致研究者们无法在先前研究的基础上建立自动化信任的研究体系.

社会心理学中的人际信任研究可以为定义自动化信任提供重要借鉴. 许多实证研究结果已经表明, 自动化信任与人际信任之间存在很多相似之处. 人对技术的反应与对他人的反应密切相关[41], 在人际信任与自动化信任之间有着很强的相似性, 特别是在复杂环境下任务完成过程中持续的信任动态之间的对应关系[15], 这种关系中的信任通常表现为受托者特征和行为的函数[42]. 神经学研究也已证明, 人际信任与人对技术的信任使用一些相同的神经机制[43−44], 其相似之处的一个潜在原因是, 在某种程度上, 人们对技术系统的信任代表了他们对这些系统的设计者的信任[17].

Billings 等回顾了 302 个信任定义, 包括 220 个人际信任定义以及 82 个自动化信任定义, 发现大量的自动化信任定义具体涉及期望、信心、风险、脆弱性、依赖、态度及合作等特征[45], 如图 2 所示, 其中横坐标表示重要特征, 纵坐标表示涉及该重要特征的自动化信任定义数量.

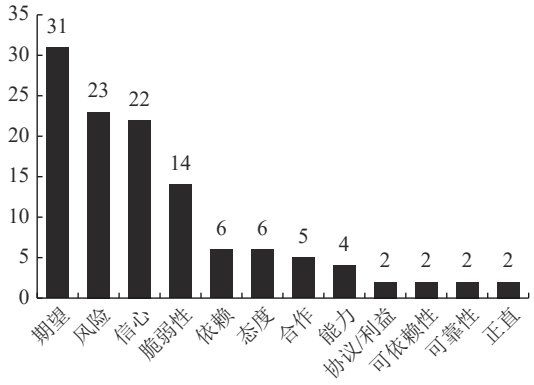


图 2　自动化信任定义涉及的重要特征

Fig. 2　Important characteristics involved in the definitions of trust in automation

大量的信任定义揭示了在人与自动化的合作关系中完成某项任务时所需的自动化信任的核心特征. 首先, 必须有一个委托者 (操作人员) 来给予信任, 必须有一个受托者 (自动化) 来接受信任, 必须有一些事情处在危险之中; 其次, 受托者必须具有某种执行任务的动机, 在人与自动化机器进行协同

作业时, 动机通常是基于设计者对机器的预期用途; 最后, 受托者必须有可能无法完成任务, 带来不确定性和风险. Lee 和 See 通过全面考察信任对依赖自动化的影响, 提出了一个包含信任核心特征的自动化信任定义, 即在以不确定性和脆弱性为特征的情况下, 代理将帮助实现个体目标的态度[27], 它是迄今为止在自动化信任研究中使用最为广泛的自动化信任定义.

1.2 自动化信任模型

在过去的几十年中, 研究人员已经建立了许多自动化信任的定性和定量模型. 然而, 信任在人机交互中的普遍存在似乎掩盖了这样一个事实 — 自动化信任的发展机制以及量化计算并没有得到充分研究. 在本文考察的文章范围内, 虽然近年来自动化信任模型相关研究在持续增长, 尤其是计算模型, 但总体文献数量仍然较少, 如图 3 所示, 其中横坐标表示文献数量, 纵坐标表示文献发表时间.

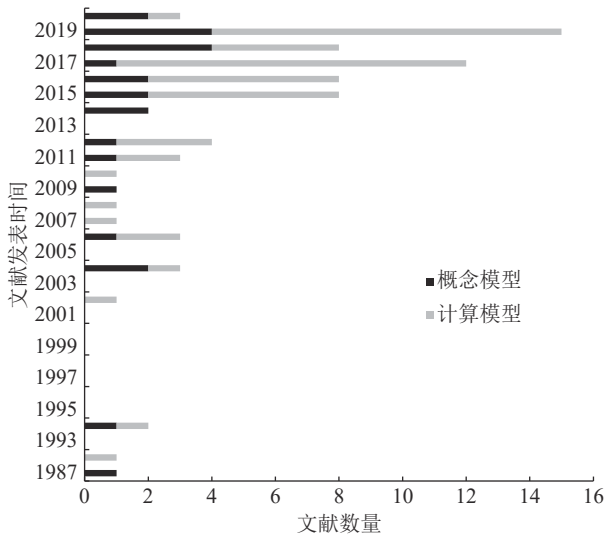


图 3 自动化信任模型文献的时间分布

Fig. 3 Time distribution of the literature on models of trust in automation

1.2.1 自动化信任的定性概念模型

准确描述自动化信任的动态发展过程是正确研究自动化信任对人机协同控制效能影响的前提. 研究人员已经建立了许多自动化信任的定性概念模型来描述其动态发展过程, 捕获可能影响该过程的重要变量. 这类模型的范围很广, 从被认为或已知有因果关系而影响自动化信任的因素列表, 到表示自动化信任与操作者对自动化行为的预测之间的因果关系的概念图. 这里将按照时间顺序选取其中为该领域的发展做出重要贡献的模型进行阐述.

1994 年, Muir 将 Barber 对人际信任的定义[46] 和 Rempel 等的信任发展模型[47] 扩展到人机关系中, 明确了在复杂的、层次化的监督控制环境中自动化信任的内涵和动态性质, 开发了一个研究自动化信任的综合框架[14]. 为了检查操作者的信任校准状态, Muir 等在该框架的基础上提出了描述自动化、操作人员的信任和对自动化行为的预测之间关系的定性模型. 随后, Muir 等进行了过程控制仿真中信任与人工干预的实验研究, 发现自动化信任与使用之间高度正相关[48]. Muir 等提出的概念模型是自动化信任研究领域的一个里程碑, 它为计划、解释和整合自动化信任的研究提供了一个理论框架, 但是随着后续研究的不断深入, 该模型已经被 Lee 和 See 提出的管理信任及其对依赖影响的动态过程的模型、Hancock 等提出的自动化信任的三因素模型以及 Hoff 和 Bashir 的三层自动化信任模型等重要模型所取代, 新的模型可以更好地概述操作者与更加复杂的自动化系统协同作业时的自动化信任.

2004 年, Lee 和 See[27] 基于理性行动理论[49] 考虑了影响信任的因素以及信任在调节对自动化的依赖中的角色, 对 Dzindolet 等的预测自动化使用的框架[50] 进行了补充, 提出了管理信任及其对依赖影响的动态过程的概念模型. 该模型指出, 自动化信任及其对行为的影响取决于操作者、环境、自动化和界面之间的动态交互. 环境因素影响信念, 而信任态度就是从信念演变而来的; 信任与其他态度 (例如, 主观工作负荷) 相结合, 形成依赖自动化的意图; 一旦意图形成, 时间限制或配置错误等因素以及自动化性能可能会影响人是否会将依赖的意图转化为行为, 进而使用自动化; 在使用自动化之后, 通过自动化显示来评估有关任务绩效的信息, 并反馈到关于自动化的信念之中. 该模型是自动化信任领域影响最为深远的模型之一, 在此之后的自动化信任研究几乎都达成了这样一个共识: 自动化信任取决于操作者、自动化和环境因素之间的动态交互. 然而, 由于信念、态度和意图之间的区别很难在实验环境中确定, 所以该模型并没有被广泛用作实证研究的基础.

2011 年, Hancock 等在人–机器人信任模型[30] 的基础上, 通过回顾自动化信任相关文献, 发展出了自动化信任的三因素模型[40]. 该模型将影响自动化信任的因素分为与人、自动化和环境相关这三种类型, 并进行了进一步的详细分类. 将与人相关的因素进一步分类为操作者特质、操作者状态、认知因素和情感因素这四种类型; 将与自动化相关的因素进一步分类为自动化特性和自动化能力; 将与环境相关的因素分类为与任务相关和与团队相关.

Schaefer 等随后又对该模型进行了大量的修改, 三个主要因素被保留, 但具体的调节因素发生了变化, 信任的前因发生了重组[35]. Hancock 等最初构建自动化信任三因素模型的目的是通过考虑自动化信任的实证研究成果, 进一步增强对人–机器人信任的理解, 但该模型已经得到更加广泛的应用.

2015 年, Hoff 和 Bashir 通过分析近年来自动化信任影响因素的实证研究, 提出了一个综合已有知识的三层信任模型[36]. 由于自动化信任变化的三个来源分别为操作者、自动化和环境, 因此, 他们将自动化信任的复杂性归结为三个层次: 倾向信任 (Dispositional trust)、情境信任 (Situational trust) 和习得信任 (Learned trust). 倾向信任是个体信任自动化的持久倾向; 情境信任依赖于交互的特定情境; 而习得信任则是基于与特定自动化系统相关的过去经验, 存在两种类型的习得信任: 初始习得信任和动态习得信任. 决定倾向信任、情境信任和初始习得信任的因素决定了初始自动化信任水平, 而动态习得信任则表示随着操作者与自动化交互的继续其自动化信任水平的后续变化. 虽然影响每一层信任的因素各不相同, 但这三层信任是相互依赖的. 环境对情境信任有很强的影响, 但操作者心理状态的情境依赖变化也会改变信任; 习得信任与情境信任密切相关, 因为它们都是由过去的经验所导致的, 所不同的是导致信任的过去经验是与自动化系统相关 (习得信任) 还是与环境相关 (情境信任). 该模型可能是目前最全面的自动化信任概念模型, 它适用于一系列自动化系统和情况, 为未来的自动化信任研究提供了一个非常有用的框架.

目前, 自动化信任概念模型正在从一般模型向针对性模型发展. 在上述较为全面且影响深远的概念模型的基础上, 一些研究者已经提出了建模自动化信任特殊方面的自动化信任概念模型[51−53] 以及与特定类型自动化相关的自动化信任概念模型[54].

1.2.2 自动化信任的定量计算模型

理解自动化信任本身并不是自动化设计人员的最终目标, 他们的最终目的是改进人机系统设计从而消除其对人机协同控制效能的负面影响. 因此, 研究人员已经提出了许多可以预测自动化信任水平或校准状态的定量计算模型来为自动化设计及部署阶段的改进工作提供指导依据.

自动化信任的计算模型可以根据不同的维度进行分类, 如概率性[55−57] 和确定性[58−59] 模型、认知[60−61] 和神经[62] 模型等. 从解决自动化开发各个阶段的自动化信任问题的角度出发, 本文根据用于生成预测的输入数据种类将自动化信任计算模型分为两种类型: 离线模型 (Offline models) 和在线模型 (Online models)[63], 自动化信任计算模型总结如表 1 所示.

表 1 自动化信任计算模型总结

Table 1 Summary of computational models of trust in automation

类型	离线信任模型	在线信任模型
输入	先验参数	先验参数及实时行为和生理及神经数据
作用	在可能的情景范围内进行模拟以预测自动化信任水平	在系统实际运行期间实时估计自动化信任水平
应用	用于自动化系统设计阶段	用于自动化系统部署阶段
结果	静态改进自动化设计	动态调整自动化行为

1) 离线模型

离线信任模型使用一组先验设置的参数作为输入来生成预测. 这类模型通常基于反馈循环, 在给定时刻根据系统的变量值确定下一时刻的人机系统状态, 包括信任水平、对自动化的依赖程度以及任务绩效等. 研究者们已经提出了许多离线信任模型.

许多早期模型都属于离线模型, 例如, Lee 等使用时间序列分析方法来建模自动化系统故障对操作者信任动态的影响[64], 该模型将故障发生情况以及自动化和人的绩效作为输入, 通过构建信任传递函数来预测操作者的自动化信任水平, 其预测结果表现为手动控制和自动控制的选择; Gao 等扩展了决策场理论来描述在监督控制情况下操作者依赖自动化的多重连续决策, 建立了描述自动化信任和依赖行为之间关系的模型[58], 该模型将操作者的自动化信任和自信的初始水平以及自动化和操作者的能力作为输入, 信任和自信分别根据随时间感知的自动化和人的绩效而变化, 利用信任和自信之间的差距来估计依赖于自动化的决策, 依赖的决策决定了下一时刻任务是由操作者手动完成还是由自动化系统自动完成.

然而, 早期离线模型对自动化信任动态的描述并不全面, 例如, 自动控制和手动控制的简单描述并不适用于许多复杂自动化系统. 最近, 一些研究者提出了更加全面的离线信任模型. 例如, Akash 等在现有的心理学文献的基础上确定了信任与经验直接相关, 并利用灰箱系统辨识技术基于 500 多名被试的行为数据建立了较为简单的、适用于反馈控制系统的三阶线性自动化信任模型[65], 该模型将经验和期望偏差作为输入, 将信任及累积信任水平作为模型输出. 它可以很好地捕获自动化信任的复杂动态, 描述不同人口统计特征之间信任行为的差异. 随后, Hu 等在此模型的基础上引入更多参数构建了二阶线性自动化信任模型, 并使用大量人类受试

者数据对模型进行参数化, 该模型可以更准确地获得自动化信任动态[66].

自动化信任的离线计算模型适用于评估人机系统在不同初始条件下的绩效趋势和总体绩效, 帮助研究人员获得对不同因素如何相互作用以及决定人类行为的深入了解. 由于离线信任模型能够仅基于一组初始参数生成预测, 它们可以用于在自动化系统尚未投入运行时预测被建模系统的行为, 因此它们自然地适合被用于自动化系统开发的设计阶段.

2) 在线模型

除了使用一些先验设置的参数值之外, 在线模型还利用系统运行过程中观察到的数据生成基于情境证据的预测, 因此, 它们可以用于实时估计信任水平. 事实上, 大多数现有的自动化信任计算模型都属于在线模型.

一些研究者根据操作者行为数据来构建在线信任模型. 例如, Xu 等建立了在线概率信任推理模型[55], 该模型基于行为数据和任务绩效使用动态贝叶斯网络来推断操作者的实时信任状态, 并且模型可以针对各个操作者进行训练, 从而对操作者的行为和态度进行可解释和个性化的描述, 与以往模型相比, 它在预测精度和响应能力上都有很大进步; Akash 等在操作者与自动化决策辅助系统交互的背景下, 建立了一个部分可观察的马尔科夫决策过程模型来描述自动化信任和工作负荷的动态变化[57], 利用被试者数据来估计模型的转换和观察概率, 研究系统透明度和操作者的经验对自动化信任和工作负荷的影响, 随后, 他们将模型的直观奖励函数集成到研究框架中, 用于评估自动化信任和工作负荷[67], 该模型可以帮助操作者制定出近乎最优的控制策略, 通过改变自动化系统透明度来实现人机协作绩效的改善.

另外一些研究者则使用操作者生理及神经数据来构建在线信任模型. 例如, Hu 等采用五种机器学习分类算法将连续的脑电图和皮肤电反应数据分类到不同的自动化信任水平, 推导出多个自动化信任模型[68], 这些模型的平均准确率为 71.57 %, 证明了心理生理学测量可以实时感知自动化信任水平. 随后, 在此基础上, Akash 等还进一步提出了另外两种方法来构建基于分类算法的自动化信任传感器模型[69]: 第一种方法是考虑一组常见的心理生理特征作为输入变量, 并使用该特征集训练得到一个通用自动化信任传感器模型; 第二种方法是考虑为每个个体定制一个特征集, 并使用该特征集训练得到一个个性化自动化信任传感器模型. 虽然使用个性化模型测量自动化信任水平的绩效优于通用模型, 但训练个性化模型需要更长的时间, 因此, 这两种方法的选择需要权衡模型的训练时间和绩效.

最近, 有学者同时使用操作者的行为数据和生理及神经数据构建在线信任模型, 并取得了更高的准确率. 例如, Akash 等在使用生理及神经数据构建的在线信任模型的基础上, 提出了一种自适应概率分类算法[70], 该算法使用马尔科夫决策过程来模拟先验概率, 采用自适应贝叶斯二次判别分类器模拟条件概率, 以脑电和行为数据为基础, 实现了自动化信任的实时测量, 并证明了模型的有效性, 该分类算法的准确率明显高于其他未考虑过程时间动态的分类算法.

自动化信任的在线计算模型使用可用的实时数据来提供信任水平的估计, 该模型可用于自动化系统部署阶段, 依据实时信任结果通过调整自动化行为、自动化透明度以及自动化等级来改善操作者行为从而提高人机协同控制任务绩效.

1.3 自动化信任影响因素

了解自动化信任影响因素对于预防不适当的信任校准是十分必要的. 在人机协同控制过程中, 不同的操作者通常具有不同水平的自动化信任[71], 这是因为信任校准过程受到除自动化本身之外的另外两个来源特性影响的结果. 大量的自动化信任实证研究揭示了自动化信任变化性的三个来源: 操作者、自动化和环境[27, 36]. 因此, 存在三种类型的自动化信任影响因素: 操作者因素、自动化因素和环境因素. 自动化信任影响因素总结如图 4 所示.

1.3.1 自动化因素

在自动化信任的研究中, 研究者们很自然地将重点放在人机系统的自动化因素上, 目前已经有一系列与自动化相关的信任影响因素被确定, 它们主要分为两种类型: 与自动化能力相关和与自动化特性相关.

1) 与自动化能力相关

与自动化能力相关的自动化信任影响因素主要有可靠性、可预测性、故障等.

可靠性是指自动化系统功能的一致性. 大量的证据表明, 在各种任务环境下, 高度可靠的自动化系统会促进操作者信任的增加[72−75], 例如, 始终表现良好的自动化比表现不佳的自动化更容易被信任[76−78]. 然而, 自动化可靠性的增加可能会导致操作者监视行为的减少[79].

可预测性是指自动化以符合操作者期望的方式执行的程度. 当操作者可以依据使用自动化的经验来预测自动化的表现时, 他会持续信任自动化[14];

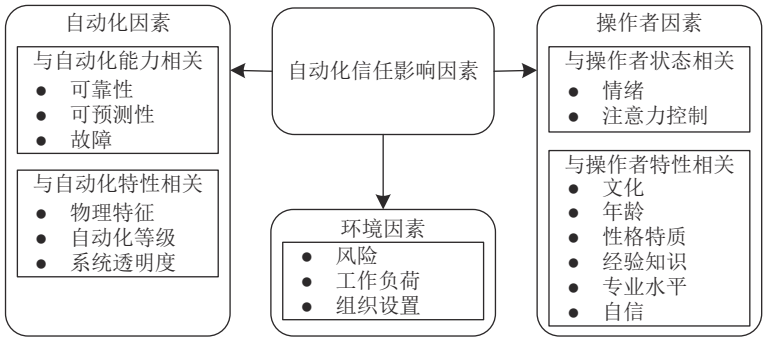


图 4　自动化信任影响因素总结

Fig. 4　Summary of factors influencing trust in automation

当自动化出现操作者意料之外的反应时, 操作者的信任水平可能会迅速下降, 这通常会导致操作者对自动化提供的信息的不使用或忽视[51].

故障是特定的系统事件, 与自动化的整体可靠性无关. 一般来说, 系统故障对自动化信任有负面影响, 故障发生通常会导致信任水平急剧下降, 即使在故障恢复后系统的表现是持续可靠的, 信任的恢复也比较慢[76, 80]. 系统故障对信任的影响也大于系统可靠运行[81], 在系统故障之后, 信任的恢复要缓慢得多, 并且通常不会达到以前的水平[64].

2) 与自动化特性相关

与自动化特性相关的自动化信任影响因素主要有自动化的物理特征、自动化等级、系统透明度等.

自动化的物理特征如界面的拟人性会使操作者表现出更强的信任弹性[82−83]. 人们对被描述为专家系统的自动化的信任程度更高[84], 但当自动化出错时, 这种信任可能会迅速降低[51].

自动化等级可能会使自动化信任的发展和改变复杂化[85]. 自动化等级越高, 操作者就越难理解, 这可能会导致其信任水平的降低[86]. 与自动化等级较高的系统相比, 操作者可能会因为对自动化的控制程度较高而更倾向于信任自动化等级较低的系统[87]. 在某些情况下, 自适应自动化可以有效地解决涉及不同自动化等级的权衡问题[88].

透明度是指自动化行为可以被理解和预测的程度[89], 设计更加透明的自动化系统可以更好地促进适当的信任, 提高任务执行绩效. 例如, 向操作者提供自动化可靠性的信息可以促进其自动化信任校准[90]; 向操作者解释自动化故障发生的原因也可以提高其信任水平[91].

1.3.2　操作者因素

虽然目前的自动化信任影响因素研究对操作者因素的关注远不及自动化因素[30], 但自动化信任是一个以人为中心的结构, 操作者因素是最重要的自动化信任影响因素. 目前研究关注的影响信任的操作者因素主要有两种类型: 较为稳定的操作者特性以及动态的操作者状态[92]. 虽然操作者的自动化信任取决于其动态状态, 但状态的预测价值往往会受到其易变性、难测性以及由大量具有交互作用的变量引起的复杂性的限制, 因此, 现有文献中对较为稳定的操作者特性研究较多.

1) 与操作者状态相关

与操作者状态相关的自动化信任影响因素主要有情绪、注意力控制等.

情绪与信任发展之间可能存在直接的关系. 积极的情绪可以显著增加信任水平[93−94], 但可能导致过度依赖[79, 95]; 与积极情绪相比, 消极情绪的影响可能更大, 它可能会导致信任下降[96] 及随后的停用[97].

注意力控制水平通常取决于操作者的工作负荷[78], 但可能受到动机、疲劳、压力或无聊等的影响[98]. 与注意力控制水平较高的操作者相比, 水平较低的操作者可能会更加依赖自动化, 即使系统的可靠性较低[99−100].

2) 与操作者特性相关

与操作者特性相关的自动化信任影响因素主要有文化、年龄、性格特质、经验知识、专业水平、自信等.

文化对人际信任具有显著影响[101], 一些研究证实了文化也影响自动化信任[102−103], 但很少有研究表明文化对自动化信任的具体影响.

针对车辆自动化 (如驾驶员预警系统[104−105]) 和决策辅助自动化 (如药物管理系统[106]) 的研究表明, 老年人比年轻人更信任自动化. 然而, 不同年龄的操作者或用户对自动化信任的评估策略可能有所不同, 年龄对信任的具体影响可能会随着情境的不同而变化[107].

操作者的某些性格特质如内向或外向[108−109] 与其总体信任倾向高度相关[110]. 性格特质对信任倾向的影响在信任发展初期占主导地位[21]. 与信任倾向

较低的个体相比, 信任倾向高的个体更可能信任可靠的系统, 但随着自动化故障的出现, 他们的信任可能会显著下降[71]. 操作者的总体信任倾向与其对特定系统的信任是不同的[17], 性格特质对自动化信任的影响可能会随着自动化和任务的不同而变化[111].

对自动化的理解是影响信任的最强因素, 其影响大于自动化的可靠性和能力[112]. 经验知识可以促进对自动化的理解[112], 它对自动化信任的发展有着直接的影响[71, 113].

提高操作者的专业水平通常会有助于其自动化信任的提高[113]. 专业水平越高, 操作者就越不可能依赖自动化[114−115]. 然而, 较高的专业知识水平可能会削弱操作者在与高可靠系统交互时监控未预期状态的能力[79].

自信已经被证明是使用自动化的重要决定因素[116]. 自信在与控制分配相关的决策过程中发挥着重要的作用, 早期研究提出了关于自信与信任的简单关系: 当信任超过自信时, 就会使用自动控制; 当自信超过信任时, 就会使用手动控制[84, 117].

1.3.3 环境因素

环境因素通常复杂多变且大部分不可控. 虽然环境因素在一般自动化应用方面研究较多, 但与自动化信任相关的研究仍然较少. 已有研究表明, 可能影响自动化信任的环境因素主要包括风险、工作负荷及组织设置等.

风险可能是影响自动化信任的最重要的环境因素之一, 对自动化的依赖是由交互过程中固有的风险水平调节的[86]. 与低风险情况相比, 一旦信任水平降低, 操作者在高风险情况下重新使用自动化需要更长的时间[116], 然而有关系统行为的预先信息可能会改变操作者对风险的看法, 当操作者知道自动化何时以及可能会如何失败时, 他们的信任不会减少[118].

工作负荷通过影响操作者监视自动化所需的时间和注意力来影响自动化信任[119]. 已经证实, 工作负荷会影响自我报告信任和依赖行为[119], 信任和依赖之间的正相关关系也会受到工作负荷的调节, 当工作负荷很高时, 无论信任水平如何, 操作者都更依赖自动化[120].

当多个操作者共同承担监视自动化的责任时, 单个操作者的自动化信任形成过程可能不同[36]; 某个操作者或主管的意见和期望可能会影响其他操作者对自动化的态度[121].

1.4 自动化信任测量方法

除了通过定量计算模型来预测自动化信任水平, 获得信任水平的另一种途径就是借助某种手段及工具发展测量方法来量化信任. 然而, 自动化信任是一种纯粹的心理结构[14], 对自动化信任进行测量是非常困难的. 已有实证研究出现的自动化信任测量方法主要有三种: 自我报告测量、行为测量和生理及神经测量. 如图 5 所示, 在这篇文章所考查的实证研究中, 大约 64 % 的研究使用某种形式的主观自我报告量表, 大约 10 % 的研究使用行为结果来推断信任, 而剩余研究则使用生理及神经指标来测量信任. 某些实证研究同时采用多种测量方法.

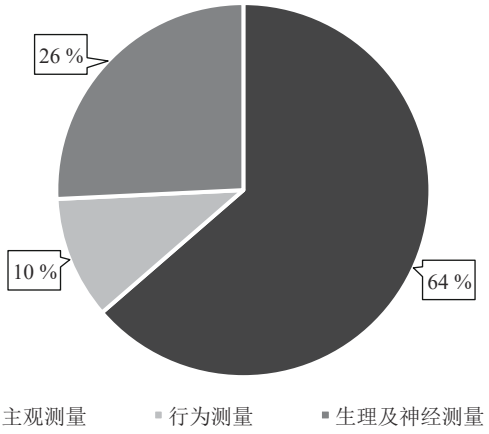


图 5 三种自动化信任测量方法的应用比例

Fig. 5 Application ratio of three trust in automation measures

1.4.1 自我报告测量

自我报告测量是唯一一种可以直接评估自动化信任水平的方法.

自动化信任的自我报告量表往往由 1~10 个量表项组成, 量表的范围通常从 “根本不 (信任)” 到 “完全 (信任)”. 量表通常采用奇数项, 这允许被试报告中立的信任水平. 到目前为止, 最常用的主观验证量表是 Jian 等开发的 7 分制 12 项量表[122], 该量表旨在衡量对自动化的总体信任程度, 具有良好的内部效度[123]. 另外, 针对不同的研究对象, 有学者开发了适用于不同研究目的自我报告量表, 例如 Mayer 等的信任倾向量表[124]、Lee 等的主观评分量表[117]、Madsen 等的人机信任量表[125]、Chien 等的跨文化自动化信任量表[126] 以及针对自动驾驶汽车[127] 和机器人[128] 等的自动化信任量表.

自我报告测量方法易于使用, 如果研究者正确构建了问卷或量表, 那么该方法可以有效地反映操作者的自动化信任水平. 然而, 该方法对交互作业具有干扰性并且难以实时捕获自动化信任的动态变化, 它在实际环境中的应用受到很大限制. 此外, 该

方法具有不可避免的缺陷, 即被试可能不能或不愿意准确报告他们的真实态度, 并且他们无法描述隐性态度对其信任水平的影响[95].

1.4.2　行为测量

为了弥补自我报告测量的缺陷, 一些研究者开始从可见的行为中来推断自动化信任水平. 使用行为度量自动化信任主要是依据遵从和依赖的概念, 即当操作员更倾向于遵从或依赖系统时, 其自动化信任水平较高, 反之则较低[129−131]. 遵从是指当自动化系统发出信号时, 操作者做出响应; 依赖则是指当自动化系统处于沉默状态或正常运行状态时, 操作者不响应[132]. 常用测量行为及其典型例子的归纳如表 2 所示.

表 2　常见的自动化信任行为测量方法总结

Table 2　Summary of common behavioural measures of trust in automation

行为	典型例子
依赖	1) 将控制权移交给自动化或从自动化收回控制权[133]. 2) 降低对自动化的监视程度[134−135].
遵从	1) 接受由自动化提供的建议或选择的动作[136]. 2) 放弃自己的决定来遵守自动化的决定[137].
其他	1) 选择手动还是使用自动化完成任务[58, 84]. 2) 选择的自动化水平[138] (操作者选择的自动化水平越高, 其信任水平越高). 3) 反应时间[139] (较长的反应时间代表较高的信任水平).

行为测量方法旨在间接评估自动化信任水平, 不具有干扰性, 并且它提供了更加一致的测量手段, 因此, 行为测量结果可以更容易地被用作建模和预测的基础. 然而, 在正常的自动化操作阶段, 某些用来推断自动化信任水平的行为可能是不可见的, 在这些情况下, 行为测量方法有一定的局限性[140]. 此外, 与自我报告测量一样, 行为测量很难捕获自动化信任的实时动态变化.

1.4.3　生理及神经测量

生理及神经测量旨在通过测量与自动化信任相关的生理及神经指标来对其进行实时测量, 虽然该方法尚处于起步阶段, 但已有文献表明它在获取自动化信任的实时动态变化方面非常有效.

目前已被证明非常具有潜力的自动化信任生理及神经测量方法主要使用眼动追踪、脑电图 (Electroencephalogram, EEG) 以及皮肤电活动 (Electrodermal activity, EDA) 等测量技术, 测量方法及其依据总结如表 3 所示.

此外, 一些研究者也探索了其他可以用于自动化信任测量的生理及神经指标或测量技术, 包括外源性催产素[136]、面部表情[141]、声音[141]、心率[141] 及功能性磁共振成像[52] 等.

生理及神经测量具有连续、实时的特点, 使用眼动追踪、EEG 及 EDA 等技术的测量方法非常具有前景, 但目前尚不清楚在使用这些方法时, 自动化信任对生理及神经指标的影响可以在多大程度上与工作负荷、压力或疲劳等其他因素产生的影响加以区分[152]. 因此, 研究者们通常将多个生理及神经指标相结合并且使用自我报告测量和行为测量方法来校准和验证生理及神经测量的结果.

2　自动化信任的研究展望

理解自动化信任并将其融入到人机系统的设计和应用中对于提高人机协同控制的绩效和安全性非常重要. 本文在自动化信任文献分析的基础上, 结合详尽的研究现状综述, 提炼出自动化信任研究趋势及现存问题, 从人机系统设计角度出发, 为未来的自动化信任研究提供一些建议.

2.1　自动化信任文献分析

国内已有一些学者开始关注信任在智能时代的人机关系中的重要作用. 北京邮电大学的刘伟认为, 人机融合智能中的一个重要课题是如何解决人与机器之间的信任问题[153]; 浙江大学的许为在其探讨以用户为中心的设计的一系列文章中指出, 智能时代

表 3　重要的生理及神经测量方法及其依据

Table 3　Important physiological and neural measures of trust in automation and their basis

测量方法	方法依据
通过眼动追踪捕获操作者的凝视行为来对自动化信任进行持续测量.	监视行为等显性行为与主观自动化信任的联系更加紧密[78]. 虽然关于自动化信任与监视行为的实验证据并不是单一的[142], 但大多数实证研究表明, 自动化信任主观评分与操作者监视频率之间存在显著的负相关关系[48]. 表征操作者监视程度的凝视行为可以为实时自动化信任测量提供可靠信息[140, 142−143].
利用 EEG 信号的图像特征来检测操作者的自动化信任状态.	许多研究检验了人际信任的神经关联[144−148], 使用神经成像工具检验自动化信任的神经关联是可行的. EEG 比其他工具 (如功能性磁共振成像) 具有更好的时间动态性[149], 在脑−机接口设计中使用 EEG 图像模式来识别用户认知和情感状态已经具有良好的准确性[149]. 自动化信任是一种认知结构, 利用 EEG 信号的图像特征来检测操作者的自动化信任校准是可行的, 并且已经取得了较高的准确性[68−69, 150].
通过 EDA 水平推断自动化信任水平	已有研究表明, 较低的自动化信任水平可能与较高的 EDA 水平相关[151]. 将该方法与其他生理及神经测量方法结合使用比单独使用某种方法的自动化信任测量准确度更高, 例如将 EDA 与眼动追踪[142] 或 EEG 结合使用[68−69].

的人机关系已经演变为人机组队式合作, 信任是人机组队合作的基本特征之一, 如何在人与自主系统之间维持适当的信任是未来人机交互的研究重点之一[154]. 然而, 国内学者目前针对自动化信任的研究非常少, 通过中国知网、维普、万方三个中文文献检索平台进行检索, 仅有两篇中文文献与自动化信任主题直接相关, 它们主要论述了自动化信任对航空安全的危害及相关改进分析[155−156].

针对上述情况, 本文选择以英文自动化信任文献为基础展开文献分析. 获得本文所综述的文献范围的具体步骤如下: 1) 确定研究主题为 “trust in automation/ automated system/ autonomous system”, 使用该主题在数据库 Web of Science 核心合集中对 1980 年 1 月至 2020 年 4 月的文献进行检索, 获得 2271 篇文献; 2) 对 2271 篇文献进行筛选, 删除重复及无关文献, 获得 415 篇文献; 3) 基于 415 篇文献所引用的全部参考文献, 对其中不在本地文献范围内的文献进行筛选, 选取以下两类文献: i) 不包含检索关键词的相关文献、ii) 与检索主题相关但不在 Web of Science 核心合集数据库中的相关文献; 4) 通过互联网英文文献检索平台 Web of Science、Google Scholar、Engineering Village 等来检索上述两类文献并添加到本地文献中, 最终获得 472 篇文献. 其中, 期刊论文 301 篇, 会议论文 171 篇, 语言为英语. 在所调查的文献范围内, 本文对自动化信任的总体发展趋势、重点应用背景及研究对象进行文献分析, 结果如图 6 所示.

2.1.1 自动化信任的总体趋势

从图 6 中的年文献发表数量统计数据可以看出, 自动化信任研究论文始现于 1987 年. 在 1987 年至 2000 年间, 自动化信任文献年发表数量较少, 共计 20 篇, 这表明在此期间的自动化信任研究处于萌芽阶段, 仅有少数研究者对其进行了零星探索. 之后的 2001 年至 2010 年间, 自动化信任年文献发表数量总体呈现上升的趋势, 但数量仍然较少, 共计 84 篇, 这表明在此期间的自动化信任研究正处于起步阶段, 发展较为缓慢. 最近十年以来, 自动化信任年发文量开始迅速增长, 2011 年至 2020 年 4 月的自动化信任研究论文总数增加到 368 篇, 自动化信任研究迎来其加速发展阶段, 自动化信任的研究范围迅速扩大, 自动化信任也已经成为许多人机交互领域学者关注的热点问题, 可以预见, 未来的自动化信任研究将保持持续快速增长的态势.

2.1.2 自动化信任的重点应用背景

自动化信任的研究涉及众多领域, 主要有军事、

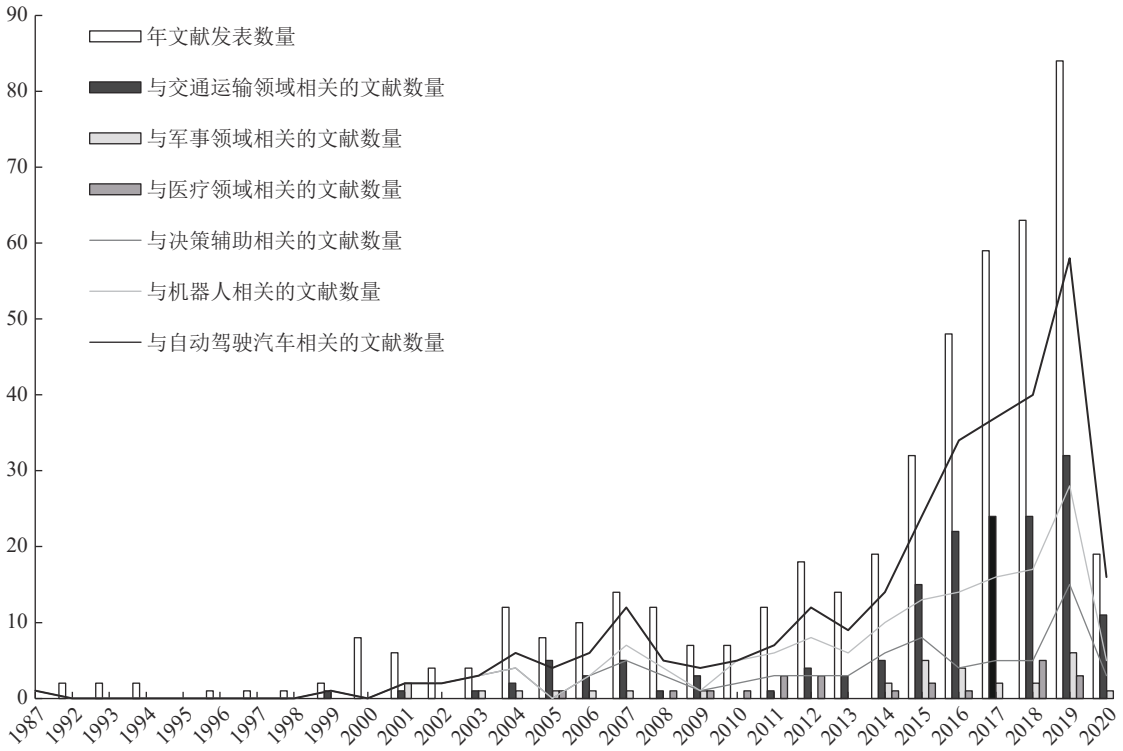


图 6 与文献发表趋势、重点应用领域及研究对象相关的自动化信任文献分析结果

Fig. 6 Results of literature analysis related to literature publication trends, key application areas and research objects of trust in automation

医疗以及交通运输等.

在军事领域, 自动化信任问题尤为突出, 因为军事环境产生了最高形式的风险、脆弱性和不确定性, 与此同时, 高风险和高节奏的情境对军事指挥和控制人员的精神和身体要求非常高, 他们经常处于极度不适和疲劳的状态, 需要高度依赖自动化系统完成团队任务[157], 错误使用自动化系统的代价可能是致命的. 因此, 随着武器装备智能化、无人化趋势日趋明显, 军事领域对自动化信任问题越来越重视.

在医疗领域, 由于辅助决策自动化系统如报警系统和建议系统被大量使用来提高决策效率, 对这些决策支持系统的不当信任很可能会导致医护人员做出错误的决策, 造成严重的医疗事故. 因此, 为了保证相关操作人员对决策支持系统保持合适的信任, 大量研究者在医疗背景下展开了与决策支持系统相关的自动化信任研究.

与交通运输相关的自动化信任研究主要集中在航空领域和汽车领域. 在航空领域, 长期以来, 飞行员、空中交通管制员或其他操作人员的自满和对自动化系统的过度依赖所导致的自动化误用一直被认为是造成航空事故的主要原因, 这些事故具有严重的经济和安全后果, 而许多实证研究已经证明, 自满和依赖与过度的自动化信任密切相关[78], 因此, 为了保证航空事业安全健康地发展, 航空领域率先开展了自动化信任相关研究. 在汽车领域, 近年来, 随着软硬件平台、人工智能和传感器技术等的进步, 自动驾驶技术得到飞速发展. 汽车制造商如特斯拉等已经制造出了商用的半自动和全自动驾驶汽车. 然而, 在全世界推广自动驾驶汽车的一个主要挑战是, 消费者对自动驾驶汽车高度不信任. 驾驶员的自动化信任对于接受和正确使用自动驾驶汽车至关重要, 因此, 以自动驾驶汽车为研究对象的自动化信任研究急速增长.

目前, 由于自动化信任相关研究仍然停留在理论阶段, 许多研究只针对特定类型的自动化系统, 而没有强调其应用背景, 与军事领域和医疗领域直接相关的自动化信任文献数量较少.

2.1.3 自动化信任的重点研究对象

对自动化信任实证研究的调查表明, 自动化信任研究对象主要有以下三类: 自动决策辅助系统、机器人和自动驾驶汽车.

自动决策辅助系统是在复杂环境中支持人类决策的自动化系统, 这些系统旨在通过提供自动生成的线索来支持人类信息分析或响应选择的认知过程, 以帮助用户正确评估给定的情况或系统状态, 并做出适当的响应. 自动决策辅助系统主要有两种功能: 报警和建议. 报警功能是简单报警系统的主要功能, 它通常被嵌入到更复杂的决策辅助系统中, 使用户了解可能需要采取行动的情况变化, 例如, 汽车导航设备为驾驶员提供驾驶建议; 飞机驾驶舱预警系统 (如交通冲突和警报系统和近地警告系统) 为飞行员提供特定的指令; 医疗专家系统提供关于病人治疗方案和药物剂量选择的建议. 自动决策辅助系统对提高人员决策效率具有重要价值, 尤其是在错误决策会对经济和安全造成严重后果的领域. 如果要实现决策辅助的好处, 就需要适当地使用它. 然而, 在实际应用中, 由于操作人员不适当的信任校准, 决策辅助系统经常被错误使用, 与自动决策辅助系统相关的自动化信任研究很有必要.

机器人经常被用于人无法到达或不安全的环境, 从事会对人造成危险的活动如行星探索、军事打击、城市搜索与救援等, 或者需要复杂技能和信息整合的活动如外科手术等. 机器人的使用正在渗透到许多不同的应用领域, 并且, 机器人与其他大多数自动化系统有着许多不同之处. 它们是可移动的, 有时它们的外观被制造为接近人类或动物, 而且它们通常被设计成在一定距离内执行动作. 这些差异可能表明, 与其他形式的自动化机器相比, 人对机器人的信任可能有所不同. 因此, 许多研究者致力于探索人对机器人的信任这种特殊类型的自动化信任.

与自动驾驶汽车相关的自动化信任研究主要分为两种类型: 1) 针对自动驾驶汽车整体; 2) 针对自适应巡航控制系统[142, 158] 和车载导航系统[159−160] 等具体子系统. 在消费者购买自动驾驶汽车之前, 针对自动驾驶汽车整体的自动化信任研究主要关注用户对自动驾驶汽车的初始信任程度对其接受程度与购买意愿的影响[148, 161−162], 以及影响初始信任的因素[123, 163]; 在用户与自动驾驶汽车的交互过程中, 用户对具体子系统的信任决定了其能否正确使用自动驾驶汽车, 研究该过程的自动化信任动态、影响因素、信任量化与测量等[164] 有助于汽车设计者改进相关子系统设计, 确保驾驶员建立适当的自动化信任, 以预期的方式正确使用自动驾驶汽车.

2.1.4 自动化信任的主要研究团体

通过文献分析可知, 共有 36 个国家及地区的 422 个研究机构发表了自动化信任研究论文. 自动化信任主要研究团体及其研究贡献总结如表 4 所示.

2.2 自动化信任的研究展望

自动化信任对于解决人与自动化机器团队合作中的模糊性和不确定性问题, 从而最大限度地发挥

表 4 自动化信任的主要研究团体及其研究贡献

Table 4 Main research groups of trust in automation and their research contributions

序号	国别	机构	团队及代表学者	研究贡献	文献数
1	美国	美国陆军研究实验室	人类研究和工程局的 Chen	提出基于系统透明度的一系列自动化信任校准方法	26
2	美国	美国空军研究实验室	人类信任与交互分部的 Lyons	进行军事背景下的自动化信任应用研究	24
3	美国	中佛罗里达大学	仿真模拟与培训学院的 Hancock	建立人–机器人信任的理论体系并进行相关影响因素实证研究	21
4	美国	克莱姆森大学	机械工程系的 Saeidi 和 Wang	建立基于信任计算模型的自主分配策略来提高人机协作效能	20
5	美国	乔治梅森大学	心理学系的 de Visser	建立并完善自动化信任修复相关理论, 着重研究自动化的拟人特征对信任修复的作用	18
6	日本	筑波大学	风险工程系的 Itoh 和 Inagaki	基于自动化信任校准的人–自动驾驶汽车协同系统设计方法	14

人与自动化机器各自的优势具有直接意义, 在人机系统设计中充分考虑自动化信任的影响是实现更加安全有效的人机协同控制的关键. 近年来, 自动化信任相关研究数量快速增长, 研究范围迅速扩大, 研究内容也越来越多地涉及众多应用背景下的多种自动化对象及用户, 自动化信任已经成为复杂人机系统设计中的焦点问题. 然而, 目前的自动化信任研究仍然处于理论研究阶段, 并没有被用于解决人机系统设计中存在的实际问题, 确保操作者保持合适自动化信任校准状态的人机系统设计方法体系还处于起步阶段, 这主要是由于自动化信任相关理论及实证研究尚不充分. 本文在详尽的研究综述及文献分析的基础上, 从人机系统设计的角度出发, 提炼出现有自动化信任研究存在的共性问题并给出可能的解决方案.

目前, 自动化信任研究尚不充分, 在自动化信任的理论研究、量化计算及实证研究三个方面仍然存在如下问题:

1) 在自动化信任的理论研究方面. 目前, 虽然已经出现了常用的自动化信任定义和影响深远的一般自动化信任概念模型, 但自动化信任定义仍然缺乏清晰性和一致性, 自动化信任概念模型也缺乏针对性和可操作性. 自动化信任是一种隐藏的心理结构, 研究者们通常在明确信任定义的基础上建立某些假设来推断其动态发展过程, 为后续的计算建模、影响因素及测量方法研究奠定理论基础, 模糊以及不一致的自动化信任定义会对后续研究方向产生根本性影响, 这使得学者在综合先前工作的基础上建立研究变得困难. 自动化信任对情境及自动化对象高度敏感, 一般自动化概念模型可操作性较差, 只能对某种应用背景下的、针对特定类型自动化对象的人机系统设计提供有限指导, 而具有针对性的特殊自动化信任概念模型较少且研究者们尚未达成共识.

2) 在自动化信任的量化计算方面. 目前研究主要通过在发展自动化信任测量方法的基础上建立定量计算模型对信任水平进行预测. 如前所述, 自动化信任计算模型可以分为两种类型: 用于在人机系统设计阶段评估系统绩效以确定可能的改进或干预措施的离线模型、用于在人机系统部署阶段估计实时信任状态以触发自动化机器适应行为的在线模型, 这些计算模型为开发具有信任意识的人机协同控制框架开辟了道路. 然而, 三种主要的自动化信任测量方法都存在其不足之处, 这对以此为基础的计算模型预测结果的准确性产生严重危害. 更重要的是, 现有计算模型具有高度的情境和对象依赖性, 离线模型的预测能力非常有限, 而大多数在线模型也无法满足实时准确地预测自动化信任的现实需求.

3) 在自动化信任的实证研究方面. 自动化信任的实证研究主要与探究自动化信任影响因素有关. 自动化信任的量化计算用于评估操作者信任状态确定是否改进人机系统设计或行为, 而影响因素的实证研究则提供改进人机系统以实现自动化信任校准所需调整的设计要素. 在自动化信任影响因素的实证研究中, 与自动化相关的影响因素研究占据很大比例, 但这些研究工作的重心集中在与自动化能力相关的因素上, 随着自动化能力的提高, 与自动化特性相关的因素以及与操作者相关的因素在自动化信任校准中的重要性日益突显, 而与此相关的实证研究并不充分.

针对上述问题, 未来的自动化信任研究工作可以考虑从以下方面着手:

1) 在现有研究的基础上, 结合认知心理学、脑科学以及人机交互领域的最新进展, 对自动化信任这一隐藏的心理状态的本质及其动态发展过程做出更加合理的推断, 明确自动化信任定义, 完善一般自动化信任概念模型, 并在此基础上充分考虑人机协同控制应用背景、任务情境、自动化对象以及操

作人员特点, 发展更具针对性和可操作性的自动化信任概念模型.

2) 在现有测量方法的基础上, 识别多种测量方法结果之间的不一致, 寻找行为指标和生理及神经指标与主观信任水平的对应关系, 确定更加准确的实时测量指标, 以识别动态自动化信任的基本状态(适当的信任、信任不足和过度信任). 在准确的自动化信任测量方法的基础上, 根据计算模型的预期用途, 准确表示模型作用条件, 涵盖产生重要影响的环境及个体因素, 并且考虑未建模因素对模型性能的不良影响, 构建满足不同设计阶段需求的自动化信任计算模型以改进人机系统, 使操作者达到合适的自动化信任校准状态.

3) 目前的实证研究关注构建 "透明" 的自动化系统和个性化的自动化信任校准. 操作者对自动化的理解是最为重要的自动化信任影响变量, 而提高自动化系统透明度可以显著增强理解, 因此, 研究者对系统透明度这个自动化特性的兴趣日益增加, 最近的研究特别强调了反馈、反馈的透明度与自动化信任校准之间的相互作用, 并且致力于寻找合适的透明度水平以在达到自动化信任校准的同时消除透明度提高给其他人类认知变量如工作负荷等带来的负面影响, 未来的实证研究仍然应该继续将系统透明度作为研究重点. 个性化的自动化信任校准主要用于针对用户广泛且多样化的自动化机器如自动驾驶汽车等的设计改进之中, 这使得定制化的自动化设计成为可能. 由于动态的操作者状态预测价值较差, 较为稳定的操作者特质成为个性化自动化信任校准的关键, 虽然此类影响因素如文化、年龄等已经被证明会影响自动化信任, 但具体影响效应尚不明晰, 一些实证研究甚至会得出相互矛盾的结论, 需要进一步明确与个体特质相关的因素对自动化信任的影响效应. 此外, 开发可靠有效的操作者特质测量工具如心理学量表等对实证研究工作非常重要, 测量工具应能解释与自动化信任相关的个体特征的大部分差异, 预测和突出与自动化交互的个体可能出现的重要问题, 针对性地改进人机系统设计以及制定个性化的人员培训策略, 实现更加精确的个性化自动化信任校准.

References

1 Schörnig N. Unmanned systems: The robotic revolution as a challenge for arms control. *Information Technology for Peace and Security: IT Applications and Infrastructures in Conflicts, Crises, War, and Peace*. Wiesbaden: Springer, 2019. 233−256

2 Meyer G, Beiker S. *Road Vehicle Automation*. Cham: Springer, 2019. 73−109

3 Bahrin M A K, Othman M F, Azli N H N, Talib M F. Industry 4.0: A review on industrial automation and robotic. *Jurnal Teknologi*, 2016, **78**(6−13): 137−143

4 Musen M A, Middleton B, Greenes R A. Clinical decision-support systems. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. London: Springer, 2014. 643−674

5 Janssen C P, Donker S F, Brumby D P, Kun A L. History and future of human-automation interaction. *International Journal of Human-Computer Studies*, 2019, **131**: 99−107

6 Xu Wei, Ge Lie-Zhong. Engineering psychology in the era of artificial intelligence. *Advances in Psychological Science*, 2020, **28**(9): 1409−1425
(许为, 葛列众. 智能时代的工程心理学. 心理科学进展, 2020, **28**(9): 1409−1425)

7 Parisi G I, Kemker R, Part J L, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019, **113**: 54−71

8 Grigsby S S. Artificial intelligence for advanced human-machine symbiosis. In: Proceeding of the 12th International Conference on Augmented Cognition: Intelligent Technologies. Berlin, Germany: Springer, 2018. 255−266

9 Gogoll J, Uhl M. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 2018, **74**: 97−103

10 Gunning D. Explainable artificial intelligence (XAI) [Online], available: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf, April 26, 2020

11 Endsley M R. From here to autonomy: Lessons learned from human-automation research. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 2017, **59**(1): 5−27

12 Blomqvist K. The many faces of trust. *Scandinavian Journal of Management*, 1997, **13**(3): 271−286

13 Rotter J B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 1967, **35**(4): 651−665

14 Muir B M. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 1994, **37**(11): 1905−1922

15 Lewandowsky S, Mundy M, Tan G P A. The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 2000, **6**(2): 104−123

16 Muir B M. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 1987, **27**(5-6): 527−539

17 Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 1997, **39**(2): 230−253

18 Levin S. Tesla fatal crash: 'autopilot' mode sped up car before driver killed, report finds [Online], available: https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report, June 8, 2020

19 The Tesla Team. An update on last Week′s accident [Online], available: https://www.tesla.com/en_GB/blog/update-last-week%E2%80%99s-accident, March 20, 2020

20 Mayer R C, Davis J H, Schoorman F D. An integrative model of organizational trust. *Academy of Management Review*, 1995, **20**(3): 709−734

21 McKnight D H, Cummings L L, Chervany N L. Initial trust formation in new organizational relationships. *Academy of Management Review*, 1998, **23**(3): 473−490

22 Jarvenpaa S L, Knoll K, Leidner D E. Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems*, 1998, **14**(4): 29−64

23 Siau K, Shen Z X. Building customer trust in mobile commerce. *Communications of the ACM*, 2003, **46**(4): 91−94

24 Gefen D. E-commerce: The role of familiarity and trust. *Omega*, 2000, **28**(6): 725−737

25 McKnight D H, Choudhury V, Kacmar C. Trust in e-commerce vendors: A two-stage model. In: Proceeding of the 21st International Conference on Information Systems. Brisbane, Australia: Association for Information Systems, 2000. 96−103

26 Li X, Hess T J, Valacich J S. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, 2008, **17**(1): 39−71

27 Lee J D, See K A. Trust in automation: Designing for appropriate reliance. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 2004, **46**(1): 50−80

28 Pan B, Hembrooke H, Joachim′s T, Lorigo L, Gay G, Granka L. In Google we trust: Users′ & decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 2007, **12**(3): 801−823

29 Riegelsberger J, Sasse M A, McCarthy J D. The researcher′s dilemma: Evaluating trust in computer-mediated communication. *International Journal of Human-Computer Studies*, 2003, **58**(6): 759−781

30 Hancock P A, Billings D R, Schaefer K E, Chen J Y C, de Visser E J, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2011, **53**(5): 517−527

31 Billings D R, Schaefer K E, Chen J Y C, Hancock P A. Human-robot interaction: Developing trust in robots. In: Proceeding of the 7th ACM/IEEE International Conference on Human-Robot Interaction. Boston, USA: ACM, 2012. 109−110

32 Madhavan P, Wiegmann D A. Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 2007, **8**(4): 277−301

33 Walker G H, Stanton N A, Salmon P. Trust in vehicle technology. *International Journal of Vehicle Design*, 2016, **70**(2): 157−182

34 Siau K, Wang W Y. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 2018, **31**(2): 47−53

35 Schaefer K E, Chen J Y C, Szalma J L, Hancock P A. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 2016, **58**(3): 377−400

36 Hoff K A, Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 2015, **57**(3): 407−434

37 Kaber D B. Issues in human-automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation. *Journal of Cognitive Engineering and Decision Making*, 2018, **12**(1): 7−24

38 Hancock P A. Imposing limits on autonomous systems. *Ergonomics*, 2017, **60**(2): 284−291

39 Schaefer K E. The Perception and Measurement of Human-Robot Trust [Ph. D. dissertation], University of Central Florida, USA, 2013.

40 Schaefer K E, Billings D R, Szalma J L, Adams J K, Sanders T L, Chen J Y C, et al. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Human-Robot Interaction, Technical Report ARL-TR-6984, Army Research Laboratory, USA, 2014.

41 Nass C, Fogg B J, Moon Y. Can computers be teammates? *International Journal of Human-Computer Studies*, 1996, **45**(6): 669−678

42 Madhavan P, Wiegmann D A. A new look at the dynamics of human-automation trust: Is trust in humans comparable to trust in machines? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2004, **48**(3): 581−585

43 Dimoka A. What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly*, 2010, **34**(2): 373−396

44 Riedl R, Hubert M, Kenning P. Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly*, 2010, **34**(2): 397−428

45 Billings D, Schaefer K, Llorens N, Hancock P A. What is Trust? Defining the construct across domains. In: Proceeding of the American Psychological Association Conference. Florida, USA: APA, 2012. 76−84

46 Barber B. *The Logic and Limits of Trust*. New Jersey: Rutgers University Press, 1983. 15−22

47 Rempel J K, Holmes J G, Zanna M P. Trust in close relationships. *Journal of Personality and Social Psychology*, 1985, **49**(1): 95−112

48 Muir B M, Moray N. Trust in automation. Part Ⅱ. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 1996, **39**(3): 429−460

49 Ajzen I. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 1991, **50**(2): 179−211

50 Dzindolet M T, Pierce L G, Beck H P, Dawe L A, Anderson B W. Predicting misuse and disuse of combat identification systems. *Military Psychology*, 2001, **13**(3): 147−164

51 Madhavan P, Wiegmann D A. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 2007, **49**(5): 773−785

52 Goodyear K, Parasuraman R, Chernyak S, de Visser E, Madhavan P, Deshpande G, et al. An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social Neuroscience*, 2017, **12**(5): 570−581

53 Hoffmann H, Söllner M. Incorporating behavioral trust theory into system development for ubiquitous applications. *Personal and Ubiquitous Computing*, 2014, **18**(1): 117−128

54 Ekman F, Johansson M, Sochor J. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, 2018, **48**(1): 95−101

55 Xu A Q, Dudek G. OPTIMo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In: Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI). Portland, USA: IEEE, 2015. 221−228

56 Nam C, Walker P, Lewis M, Sycara K. Predicting trust in human control of swarms via inverse reinforcement learning. In: Proceeding of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Lisbon, Portugal: IEEE, 2017. 528−533

57 Akash K, Polson K, Reid T, Jain N. Improving human-machine collaboration through transparency-based feedback-part I: Human trust and workload model. *IFAC-PapersOnLine*, 2019, **51**(34): 315−321

58 Gao J, Lee J D. Extending the decision field theory to model operators′ reliance on automation in supervisory control situations.

IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2006, **36**(5): 943−959

59 Wang Y, Shi Z, Wang C, Zhang F. Human-robot mutual trust in (Semi) autonomous underwater robots. *Cooperative Robots and Sensor Networks*. Berlin: Springer, 2014. 115−137

60 Clare A S. Modeling Real-time Human-automation Collaborative Scheduling of Unmanned Vehicles [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 2013.

61 Gao F, Clare A S, Macbeth J C, Cummings M L. Modeling the impact of operator trust on performance in multiple robot control. In: Proceeding of the AAAI Spring Symposium Series. Palo Alto, USA: AAAI, 2013. 164−169

62 Hoogendoorn M, Jaffry S W, Treur J. Cognitive and neural modeling of dynamics of trust in competitive trustees. *Cognitive Systems Research*, 2012, **14**(1): 60−83

63 Hussein A, Elsawah S, Abbass H A. Towards trust-aware human-automation interaction: An overview of the potential of computational trust models. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. Hawaii, USA: University of Hawaii, 2020. 47−57

64 Lee J, Moray N. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 1992, **35**(10): 1243−1270

65 Akash K, Hu W L, Reid T, Jain N. Dynamic modeling of trust in human-machine interactions. In: Proceedings of the 2017 American Control Conference (ACC). Seattle, USA: IEEE, 2017. 1542−1548

66 Hu W L, Akash K, Reid T, Jain N. Computational modeling of the dynamics of human trust during human-machine interactions. *IEEE Transactions on Human-Machine Systems*, 2019, **49**(6): 485−497

67 Akash K, Reid T, Jain N. Improving human-machine collaboration through transparency-based feedback-part Ⅱ: Control design and synthesis. *IFAC-PapersOnLine*, 2019, **51**(34): 322−328

68 Hu W L, Akash K, Jain N, Reid T. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, 2016, **49**(32): 48−53

69 Akash K, Hu W L, Jain N, Reid T. A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, 2018, **8**(4): Article No. 27

70 Akash K, Reid T, Jain N. Adaptive probabilistic classification of dynamic processes: A case study on human trust in automation. In: Proceedings of the 2018 Annual American Control Conference (ACC). Milwaukee, USA: IEEE, 2018. 246−251

71 Merritt S M, Ilgen D R. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2008, **50**(2): 194−210

72 Bagheri N, Jamieson G A. The impact of context-related reliability on automation failure detection and scanning behaviour. In: Proceedings of the 2004 International Conference on Systems, Man and Cybernetics. The Hague, Netherlands: IEEE, 2004. 212−217

73 Cahour B, Forzy J F. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, 2009, **47**(9): 1260−1270

74 Cummings M L, Clare A, Hart C. The role of human-automation consensus in multiple unmanned vehicle scheduling. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2010, **52**(1): 17−27

75 Kraus J M, Forster Y, Hergeth S, Baumann M. Two routes to trust calibration: effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction*, 2019, **11**(3): 1−17

76 Kelly C, Boardman M, Goillau P, Jeannot E. Guidelines for Trust in Future ATM Systems: A Literature Review, Technical Report HRS/HSP-005-GUI-01, European Organization for the Safety of Air Navigation, Belgium, 2003.

77 Riley V. A general model of mixed-initiative human-machine systems. *Proceedings of the Human Factors Society Annual Meeting*, 1989, **33**(2): 124−128

78 Parasuraman R, Manzey D H. Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2010, **52**(3): 381−410

79 Bailey N R, Scerbo M W, Freeman F G, Mikulka P J, Scott L A. Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2006, **48**(4): 693−709

80 Moray N, Hiskes D, Lee J, Muir B M. Trust and Human Intervention in Automated Systems. *Expertise and Technology: Cognition & Human-computer Cooperation*. New Jersey: L. Erlbaum Associates Inc, 1995. 183−194

81 Yu K, Berkovsky S, Taib R, Conway D, Zhou J L, Chen F. User trust dynamics: An investigation driven by differences in system performance. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. Limassol, Cyprus: ACM, 2017. 307−317

82 De Visser E J, Monfort S S, McKendrick R, Smith M A B, McKnight P E, Krueger F, et al. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 2016, **22**(3): 331−349

83 Pak R, Fink N, Price M, Bass B, Sturre L. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 2012, **55**(9): 1059−1072

84 De Vries P, Midden C, Bouwhuis D. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 2003, **58**(6): 719−735

85 Moray N, Inagaki T, Itoh M. Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, 2000, **6**(1): 44−58

86 Lewis M, Sycara K, Walker P. The role of trust in human-robot interaction. *Foundations of Trusted Autonomy*. Cham: Springer, 2018. 135−159

87 Verberne F M F, Ham J, Midden C J H. Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2012, **54**(5): 799−810

88 de Visser E, Parasuraman R. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, 2011, **5**(2): 209−231

89 Endsley M R. Situation awareness in future autonomous vehicles: Beware of the unexpected. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)}. Cham, Switzerland: Springer, 2018. 303−309

90 Wang L, Jamieson G A, Hollands J G. Trust and reliance on an automated combat identification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2009, **51**(3): 281−291

91 Dzindolet M T, Peterson S A, Pomranky R A, Pierce G L, Beck

H P. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 2003, **58**(6): 697−718

92 Davis S E. Individual Differences in Operators´ Trust in Autonomous Systems: A Review of the Literature, Technical Report DST-Group-TR-3587, Joint and Operations Analysis Division, Defence Science and Technology Group, Australia, 2019.

93 Merritt S M. Affective processes in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2011, **53**(4): 356−370

94 Stokes C K, Lyons J B, Littlejohn K, Natarian J, Case E, Speranza N. Accounting for the human in cyberspace: Effects of mood on trust in automation. In: Proceedings of the 2010 International Symposium on Collaborative Technologies and Systems. Chicago, USA: IEEE, 2010. 180−187

95 Merritt S M, Heimbaugh H, LaChapell J, Lee D. I trust it, but I don′t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2013, **55**(3): 520−534

96 Ardern - Jones J, Hughes D K, Rowe P H, Mottram D R, Green C F. Attitudes and opinions of nursing and medical staff regarding the supply and storage of medicinal products before and after the installation of a drawer-based automated stock-control system. *International Journal of Pharmacy Practice*, 2009, **17**(2): 95−99

97 Gao J, Lee J D, Zhang Y. A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics*, 2006, **36**(5): 511−526

98 Reichenbach J, Onnasch L, Manzey D. Human performance consequences of automated decision aids in states of sleep loss. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2011, **53**(6): 717−728

99 Chen J Y C, Terrence P I. Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, 2009, **52**(8): 907−920

100 Chen J Y C, Barnes M J. Supervisory control of multiple robots in dynamic tasking environments. *Ergonomics*, 2012, **55**(9): 1043−1058

101 Naef M, Fehr E, Fischbacher U, Schupp J, Wagner G G. Decomposing trust: Explaining national and ethnical trust differences. *International Journal of Psychology*, 2008, **43**(3-4): 577−577

102 Huerta E, Glandon T A, Petrides Y. Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems*, 2012, **13**(4): 316−333

103 Chien S Y, Lewis M, Sycara K, Liu J S, Kumru A. Influence of cultural factors in dynamic trust in automation. In: Proceedings of the 2016 International Conference on Systems, Man, and Cybernetics (SMC). Budapest, Hungary: IEEE, 2016. 2884−2889

104 Donmez B, Boyle L N, Lee J D, McGehee D V. Drivers′ attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2006, **9**(6): 387−398

105 Kircher K, Thorslund B. Effects of road surface appearance and low friction warning systems on driver behaviour and confidence in the warning system. *Ergonomics*, 2009, **52**(2): 165−176

106 Ho G, Wheatley D, Scialfa C T. Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 2005, **17**(6): 690−710

107 Steinke F, Fritsch T, Silbermann L. Trust in ambient assisted living (AAL) − a systematic review of trust in automation and assistance systems. *International Journal on Advances in Life Sciences*, 2012, **4**(3-4): 77−88

108 McBride M, Morgan S. Trust calibration for automated decision aids [Online], available: https://www.researchgate.net/publication/303168234_Trust_calibration_for_automated_decision_aids, May 15, 2020

109 Gaines Jr S O, Panter A T, Lyde M D, Steers W N, Rusbult C E, Cox C L, et al. Evaluating the circumplexity of interpersonal traits and the manifestation of interpersonal traits in interpersonal trust. *Journal of Personality and Social Psychology*, 1997, **73**(3): 610−623

110 Looije R, Neerincx M A, Cnossen F. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, 2010, **68**(6): 386−397

111 Szalma J L, Taylor G S. Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, 2011, **17**(2): 71−96

112 Balfe N, Sharples S, Wilson J R. Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2018, **60**(4): 477−495

113 Rajaonah B, Anceaux F, Vienne F. Trust and the use of adaptive cruise control: A study of a cut-in situation. *Cognition, Technology & Work*, 2006, **8**(2): 146−155

114 Fan X C, Oh S, McNeese M, Yen J, Cuevas H, Strater L, et al. The influence of agent reliability on trust in human-agent collaboration. In: Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction. Funchal, Portugal: Association for Computing Machinery, 2008. Article No: 7

115 Sanchez J, Rogers W A, Fisk A D, Rovira E. Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, 2014, **15**(2): 134−160

116 Riley V. Operator reliance on automation: Theory and data. *Automation and Human Performance: Theory and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996. 19−35

117 Lee J D, Moray N. Trust, self-confidence, and operators′ adaptation to automation. *International Journal of Human-Computer Studies*, 1994, **40**(1): 153−184

118 Perkins L A, Miller J E, Hashemi A, Burns G. Designing for human-centered systems: Situational risk as a factor of trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2010, **54**(25): 2130−2134

119 Bindewald J M, Rusnock C F, Miller M E. Measuring human trust behavior in human-machine teams. In: Proceedings of the AHFE 2017 International Conference on Applied Human Factors in Simulation and Modeling. Los Angeles, USA: Springer, 2017. 47−58

120 Biros D P, Daly M, Gunsch G. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, 2004, **13**(2): 173−189

121 Workman M. Expert decision support system use, disuse, and misuse: A study using the theory of planned behavior. *Computers in Human Behavior*, 2005, **21**(2): 211−231

122 Jian J Y, Bisantz A M, Drury C G. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 2000, **4**(1): 53−71

123 Buckley L, Kaye S A, Pradhan A K. Psychosocial factors associated with intended use of automated vehicles: A simulated driving study. *Accident Analysis & Prevention*, 2018, **115**: 202−208

124 Mayer R C, Davis J H. The effect of the performance appraisal

system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, 1999, **84**(1): 123−136

125 Madsen M, Gregor S. Measuring human-computer trust. In: Proceedings of the 11th Australasian Conference on Information Systems. Brisbane, Australia: Australasian Association for Information Systems, 2000. 6−8

126 Chien S Y, Semnani-Azad Z, Lewis M, Sycara K. Towards the development of an inter-cultural scale to measure trust in automation. In: Proceedings of the 6th International Conference on Cross-cultural Design. Heraklion, Greece: Springer, 2014. 35−36

127 Garcia D, Kreutzer C, Badillo-Urquiola K, Mouloua M. Measuring trust of autonomous vehicles: A development and validation study. In: Proceedings of the 2015 International Conference on Human-Computer Interaction. Los Angeles, USA: Springer, 2015. 610−615

128 Yagoda R E, Gillan D J. You want me to trust a ROBOT? The development of a human-robot interaction trust scale. *International Journal of Social Robotics*, 2012, **4**(3): 235−248

129 Dixon S R, Wickens C D. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2006, **48**(3): 474−486

130 Chiou E, Lee J D. Beyond reliance and compliance: Human-automation coordination and cooperation. In: Proceedings of the 59th International Annual Meeting of the Human Factors and Ergonomics Society. Los Angeles, USA: SAGE, 2015. 159−199

131 Bindewald J M, Rusnock C F, Miller M E. Measuring human trust behavior in human-machine teams. In: Proceedings of the AHFE 2017 International Conference on Applied Human Factors in Simulation and Modeling. Los Angeles, USA: Springer, 2017. 47−58

132 Chancey E T, Bliss J P, Yamani Y, Handley H A H. Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2017, **59**(3): 333−345

133 Gremillion G M, Metcalfe J S, Marathe A R, Paul V J, Christensen J, Drnec K, et al. Analysis of trust in autonomy for convoy operations. In: Proceedings of SPIE 9836, Micro-and Nanotechnology Sensors, Systems, and Applications VIII. Washington, USA: SPIE, 2016. 9836−9838

134 Basu C, Singhal M. Trust dynamics in human autonomous vehicle interaction: A review of trust models. In: Proceedings of the 2016 AAAI Spring Symposium Series. Palo Alto, USA: AAAI, 2016. 238−245

135 Hester M, Lee K, Dyre B P. "Driver Take Over": A preliminary exploration of driver trust and performance in autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2017, **61**(1): 1969−1973

136 De Visser E J, Monfort S S, Goodyear K, Lu L, O'Hara M, Lee M R, et al. A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2017, **59**(1): 116−133

137 Gaudiello I, Zibetti E, Lefort S, Chetouani M, Ivaldi S. Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior*, 2016, **61**: 633−655

138 Desai M, Kaniarasu P, Medvedev M, Steinfeld A, Yanco H. Impact of robot failures and feedback on real-time trust. In: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Tokyo, Japan: IEEE, 2013. 251−258

139 Payre W, Cestac J, Delhomme P. Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2016, **58**(2): 229−241

140 Hergeth S, Lorenz L, Vilimek R, Krems J F. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2016, **58**(3): 509−519

141 Khalid H M, Shiung L W, Nooralishahi P, Rasool Z, Helander M G, Kiong L C, Ai-Vyrn C. Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2016, **60**(1): 697−701

142 Gold C, Körber M, Hohenberger C, Lechner D, Bengler K. Trust in automation-Before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, 2015, **3**: 3025−3032

143 Walker F, Verwey W B, Martens M. Gaze behaviour as a measure of trust in automated vehicles. In: Proceedings of the 6th Humanist Conference. Washington, USA: HUMANIST, 2018. 117−123

144 Adolphs R. Trust in the brain. *Nature Neuroscience*, 2002, **5**(3): 192−193

145 Delgado M R, Frank R H, Phelps E A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 2005, **8**(11): 1611−1618

146 King-Casas B, Tomlin D, Anen C, Camerer C F, Quartz S R, Montague P R. Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, 2005, **308**(5718): 78−83

147 Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, et al. Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(50): 20084−20089

148 Long Y, Jiang X, Zhou X. To believe or not to believe: Trust choice modulates brain responses in outcome evaluation. *Neuroscience*, 2012, **200**: 50−58

149 Minguillon J, Lopez-Gordo M A, Pelayo F. Trends in EEG-BCI for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, 2017, **31**: 407−418

150 Choo S, Sanders N, Kim N, Kim W, Nam C S, Fitts E P. Detecting human trust calibration in automation: A deep learning approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2019, **63**(1): 88−90

151 Morris D M, Erno J M, Pilcher J J. Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2017, **61**(1): 1759−1762

152 Drnec K, Marathe A R, Lukos J R, Metcalfe J S. From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Frontiers in Human Neuroscience*, 2016, **10**: Article No. 290

153 Liu Wei. Current situation and prospect of human-computer fusion intelligence. *Governance*, 2019, (4): 7−15
(刘伟. 人机融合智能的现状与展望. 国家治理, 2019, (4): 7−15)

154 Xu Wei. User-centered design (V): From automation to the autonomy and autonomous vehicles in the intelligence era. *Chinese Journal of Applied Psychology*, 2020, **26**(2): 108−128
(许为. 五论以用户为中心的设计: 从自动化到智能时代的自主化以及自动驾驶车. 应用心理学, 2020, **26**(2): 108−128)

155 Wang Xin-Ye, Li Yuan, Chang Ming, You Xu-Qun. The detriments and improvement of automation trust and dependence to aviation safety. *Advances in Psychological Science*, 2017, **25**(9):

1614−1622
(王新野, 李苑, 常明, 游旭群. 自动化信任和依赖对航空安全的危害及其改进. 心理科学进展, 2017, **25**(9): 1614−1622)

156 Cao Qing-Long. Analysis of the detriments and improvement of automation trust and dependence to aviation safety. *Technology and Market*, 2018, **25**(4): 160
(曹清龙. 自动化信任和依赖对航空安全的危害及其改进分析. 技术与市场, 2018, **25**(4): 160)

157 Adams B D, Webb R D G. Trust in small military teams. In: Proceedings of the 7th International Command and Control Technology Symposium. Virginia, USA: DTIC, 2002. 1−20

158 Beggiato M, Pereira M, Petzoldt T, Krems J. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2015, **35**: 75−84

159 Reimer B. Driver assistance systems and the transition to automated vehicles: A path to increase older adult safety and mobility? *Public Policy & Aging Report*, 2014, **24**(1): 27−31

160 Large D R, Burnett G E. The effect of different navigation voices on trust and attention while using in-vehicle navigation systems. *Journal of Safety Research*, 2014, **49**: 69.e1−75

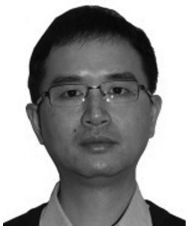
161 Zhang T R, Tao D, Qu X D, Zhang X Y, Zeng J H, Zhu H Y, et al. Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transportation Research Part C: Emerging Technologies*, 2020, **112**: 220−233

162 Choi J K, Ji Y G. Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 2015, **31**(10): 692−702

163 Kaur K, Rampersad G. Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, 2018, **48**: 87−96

164 Lee J D, Kolodge K. Exploring trust in self-driving vehicles through text analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2020, **62**(2): 260−277

董文莉 北京交通大学电子信息工程学院博士研究生. 2017 年获得郑州大学轨道交通信号与控制学士学位. 主要研究方向为自动化信任和计算认知建模.
E-mail: wldong_bjtu@163.com
(**DONG Wen-Li** Ph. D. candidate at the School of Electronic and Information Engineering, Beijing Jiaotong University. She received her bachelor degree in Rail Transportation Signaling and Control from Zhengzhou University in 2017. Her research interest covers trust in automation and computational cognitive modeling.)



方卫宁 北京交通大学轨道交通控制与安全国家重点实验室教授. 1996 年获得重庆大学博士学位. 主要研究方向为人因工程 , 轨道交通安全模拟与仿真. 本文通信作者.
E-mail: wnfang@bjtu.edu.cn
(**FANG Wei-Ning** Professor at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He received his Ph. D. degree from Chongqing University in 1996. His research interest covers ergonomics, intelligent transportation systems, system reliability and safety, and railway simulation. Corresponding author of this paper.)