Advances in Philosophy 哲学进展, 2026, 15(5), 45-51
Published Online May 2026 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2026.155199

Hans汉斯

“向善”且“为善”：多模态情感识别的伦理约束研究

陈郭蓉

扬州大学社会发展学院，江苏 扬州

收稿日期：2026年4月6日；录用日期：2026年4月26日；发布日期：2026年5月11日

摘　要

多模态情感识别作为人工智能情感计算领域的前沿方向，其技术融合了计算机视觉、语音处理、生理信号分析等多种模态，显著提升了情感识别的准确性与鲁棒性。然而，技术能力的跃升也带来了伦理风险的质变——从传统AI的“行为观察”迈向对个体内在情感状态的深度识别，对精神隐私、情感自主性乃至人类主体性构成深层挑战。本研究系统剖析了多模态情感识别的技术特性与伦理风险，指出其具有学科交叉复杂性、伦理碰撞冲突性、善恶影响非对称性等新特点，并在数据隐私、算法歧视、情感操控、人性异化四个维度展开风险分析。在此基础上，基于《关于加强科技伦理治理的意见》与《人工智能伦理治理标准化指南》，建构了“5 + 10 + X”的伦理风险评价准则体系。进一步引入海德格尔“共在”思想、梅洛－庞蒂身体现象学与马克思异化理论，尝试构建“哲学 + 技术”深度融合的治理框架，为多模态情感识别的伦理约束与制度规范提供理论支撑与实践路径。

关键词

多模态情感识别，精神隐私，科技伦理

“Striving for Good and Acting for Good”: A Study on Ethical Constraints in Multimodal Emotion Recognition

Guorong Chen

College of Social Development, Yangzhou University, Yangzhou Jiangsu

Received: April 6, 2026; accepted: April 26, 2026; published: May 11, 2026

Abstract

As a cutting-edge direction in the field of affective computing within artificial intelligence, multimodal

文章引用: 陈郭蓉. “向善”且“为善”: 多模态情感识别的伦理约束研究[J]. 哲学进展, 2026, 15(5): 45-51.
DOI: 10.12677/acpp.2026.155199

emotion recognition integrates various modalities such as computer vision, speech processing, and physiological signal analysis, significantly enhancing the accuracy and robustness of emotion recognition. However, this leap in technological capability also brings about a qualitative shift in ethical risks—moving from the "observation of behavior" characteristic of traditional AI to the "penetration of the inner self," posing profound challenges to mental privacy, emotional autonomy, and even human subjectivity. This study systematically analyzes the technological characteristics and ethical risks of multimodal emotion recognition, identifying new features such as the complexity of interdisciplinary integration, the clash of ethical conflicts, and the asymmetry between beneficial and harmful impacts. It further examines risks across four dimensions: data privacy, algorithmic discrimination, emotional manipulation, and human alienation. On this basis, drawing from the "Opinions on Strengthening Ethical Governance in Science and Technology" and the "Guidelines for Ethical Governance in Artificial Intelligence Standardization", this study constructs a "5 + 10 + X" ethical risk evaluation criteria system. Additionally, by incorporating Heidegger's concept of "Being-with" (Mitsein), Merleau-Ponty's phenomenology of the body, and Marx's theory of alienation, this study attempts to develop a governance framework that deeply integrates philosophy and technology, providing theoretical support and practical pathways for the ethical constraints and institutional regulation of multimodal emotion recognition.

Keywords

Multimodal Emotion Recognition, Mental Privacy, Ethics of Science and Technology

1. 引言

1997 年，Gunther Kress 与 Theo van Leeuwen 最先提出“多模态”(Multi-modal)这一概念，同年，MIT 媒体实验的 Rosalind Picard 教授提出“情感计算”的概念。多模态情感识别(Multimodal Emotion Recognition)作为情感计算的一个重要分支，其研究始于 20 世纪 90 年代，随着计算机视觉、语音处理和自然语言处理技术的发展，研究者发现单一模态不足以全面捕捉复杂的人类情感，因而开始探索融合面部表情、语音、文本、生理信号等多种模态的方法。以“多模态情感识别”作为明确术语和系统性研究方向则是在 2000 年至 2010 年逐步形成并得到广泛认可的。2022 年 12 月 9 日，由之江实验室发起，德勤中国、上海科学技术出版社、中国科学院文献情报中心和英国工程技术学会共同参加编写的《情感计算白皮书》(以下简称《白皮书》)面向全球正式发布。《白皮书》指出，多模态情感计算是当下人工智能领域最热门的话题之一[1]。

正如“科技向善”理念所强调的，人工智能的发展必须以增进人类福祉为根本目标，在技术进步与伦理规范之间寻求动态平衡[2]。随着多模态情感识别的广泛应用，隐私保护、算法歧视、情感操纵等问题与风险日益凸显，多种伦理挑战也随之而来，包括机器决策的道德责任、偏见与歧视、机器智能权利、责任归属、社会不平等、人的主体性的消解等。为了应对这些风险挑战，我们需要对多模态情感识别的责任边界、嵌入的伦理与价值观、人机共生伦理关系等进行研究，从而确保多模态情感识别的发展不偏离伦理的轨道，始终与人类的价值与利益保持一致。现有研究多聚焦于技术性能优化或者人工智能普遍的伦理问题治理，缺乏对多模态情感识别伦理风险的特殊性的解构和系统性的应对。

本研究从理论层面将人工智能伦理研究拓展至精神隐私领域，探讨‘精神隐私权’这一拟议中的概

念，并通过跨学科融合构建多维分析框架；在现实层面，建构“5 + 10 + X”伦理风险评价准则体系，为政策完善与行业负责任创新提供具体制度规范与实践指引。

2. 多模态情感识别的新特点

(一) 学科交叉的复杂性

多模态情感识别涉及计算机科学、信号处理、心理学、伦理学等多个研究领域[3]，其融合一方面孕育出了集工程学、心理学、伦理学、认知科学等学科于一体的智能情感计算程序，超越了人机交互领域原有的信息处理技术，同时也超越了技术价值的中立立场，带来了新兴的技术福利；另一方面，这也使得学科之间不同的逻辑思维、价值观念、方法论等相互干扰、相互矛盾。比如，工程学追求算法的局部最优解和可量化的准确指标，与心理学的情感主观性、情感依赖性和流动模糊性存在根本张力。可见，多学科在情感这一人类最私密领域的交汇具有不容忽视的复杂性。

(二) 科技伦理碰撞的冲突性

多模态情绪识别是指让计算机通过综合分析多种不同类型的数据(如语言、声音、面部表情、肢体动作等) [3]，来更准确、更全面地判断一个人当前的情感状态，基于表情、语音和语言的多模态情感识别结合多个感官通道的信息，有着强烈的科技主义思维。多模态情感识别的目的是让机器能像人类一样，全方位地感知并理解我们的情绪，提升情感判断的准确性和鲁棒性，推动人机交互向着更自然、更人性化的方向发展，也能让多模态情感识别赋能更多垂直领域，解决更多的实际问题。但同时该技术也存在一系列伦理问题，比如隐私、偏见等通常问题，更高级的伦理风险是人类的情感自主性被隐性剥夺，这对情感识别的合理性形成严峻挑战，同时也引发了一系列“机器是否会独立思考”的伦理问题[4]。

(三) 善恶两面的非对称性

区别于一般技术的“双刃剑”隐喻，多模态情感识别的特殊性在于其风险与收益往往呈现显著的非对称性分布——获益主体与风险承担主体常常相互分离。从研发过程看，算法开发者致力于追求情感识别的准确率与响应速率，由于情感数据所蕴含的极高商业价值，极易诱发对技术边界的僭越，将“能够识别”等同于“应该识别”，此种技术为开发者带来了市场先机，却把情感剥夺的风险转嫁给了被识别者；从使用结果看，技术应用中的收益与风险分离更为凸显。比如在教育场景中，校方或平台通过多模态系统监测学生专注度以提升管理效能，而学生因偶发的“情感非典型表达”被算法标记为“问题个体”进而影响发展机会。这种“开发者获利、社会承担成本”的非对称结构，使得多模态情感识别的伦理治理必须直面权力结构中的责任分配正义问题[4]。

3. 多模态情感识别的核心伦理风险

(一) 数据隐私与安全风险

多模态情感识别的数据隐私问题，已经超越了传统个人信息保护的范畴，在现行法律下，对“精神隐私”构成了双重挑战。

根据《中华人民共和国民法典》及相关法学权威解读，“精神隐私”并非独立权利，而是隐私在精神或心理层面的体现，属于精神性人格权的一部分，隐私权保护的是自然人不愿为他人知晓的私密空间、私密活动、私密信息以及私人生活安宁。在人格语境下，“精神隐私”可理解为，内心不被他人非法刺探、干预或公开的思想、情感、记忆、心理活动等和精神健康状况、心理诊疗记录、创伤经历等敏感信息；随着脑机接口、神经成像等技术发展，在新兴科技语境下，“精神隐私”也包括个体脑电波、神经信号、大脑活动数据等。

传统意义上的精神隐私侵犯，往往依赖于物理侵入或明显的非法获取手段，但多模态情感识别可以

通过远距离视频采集、可穿戴设备传感、语音情绪分析等非接触式手段，能够在个体毫无感知的情况下，将其面部微表情、语音变化、心率变异等承载情感状态的生物信号捕获并还原为情感数据，依据对精神隐私权的界定，精神隐私权主要保护的就是内心思想、情感状态、心理活动等不被非法刺探，然而，多模态识别技术的介入，使得这种“不被刺探的权利”面临瓦解。

多模态情感识别系统不仅存在敏感数据的采集问题，还存在数据存储与共享的合规性问题。多模态情感识别系统往往通过“用户协议”或“隐私条款”的一揽子勾选获取授权，但此类协议普遍存在语言晦涩、授权范围宽泛、目的描述模糊等问题，用户即便点击“同意”，也往往无法真正知晓其面部微表情、语音语调、心率变异等生物特征数据将被如何存储、与哪些第三方共享、用于何种目的，这种“点击即同意”的机制，将《民法典》所要求的“自愿、明确”的知情同意降格为形式主义的程序空转。现行多数系统的数据共享协议缺乏透明度，往往以“提升服务质量”“优化用户体验”等模糊表述掩盖数据二次利用的真实目的，现行法律框架下以“知情同意”为核心的个人信息保护机制，在多模态数据的管理中正面临着失效。

(二) 算法偏见与歧视

人类学与跨文化心理学研究表明，虽然基本情感的面部表达存在一定的跨文化共通性，但情感表达的规则、强度、语境与组合方式却高度依赖于文化规范，例如，东亚文化强调“内敛”，个体在社交场景中倾向于抑制强烈的情感外露；而西方文化则更推崇“外显”，鼓励情感的开放表达。当基于西方样本训练的情感识别模型应用于东亚人群时，后者内敛的情感表达便极易被误判为“冷漠”“平淡”甚至“消极”，多模态融合会强化这种文化偏见，很容易形成误判，这不仅会导致服务体验的偏差，更会演变为对特定文化群体的病理性标签。

如果说文化偏见是“跨空间”的差异，那么数据不足则是涉及“跨群体”的差异。多模态情感识别的建立高度依赖于训练数据的特征，出于对数据采集成本、商业市场群体等因素的考量，训练数据往往集中于年轻、健康的社会群体，这种数据的结构失衡，极易引起技术性的误判并使得社会少数群体面临资源分配的不公。例如，老年人的情感表达往往因为生理机能衰退而呈现如语音震颤、面部肌肉活动减弱等区别于年轻人的特点；残障人士的情感表达路径也同样异于常人，面部差异者甚至不能呈现出“标准”的表情。算法试图将流动、多元的情感体验纳入标准化的分类中，然而情感的个体差异性却使得多模态情感识别技术面临着根本性的认识论困境。

(三) 情感操控与人性异化

情感识别技术通过分析用户情绪弱点(如孤独、焦虑、抑郁)，精准设计诱导性服务或消费场景，例如，部分 AI 伴侣应用会在用户深夜倾诉孤独时，推荐付费语音陪伴服务或虚拟礼物，将情感需求转化为商业利润，这种“情感数据的商业化利用”模式的运作逻辑在于：技术并非被动响应用户需求，而是主动识别情绪脆弱窗口，在用户心理防御最薄弱的时刻切入商业诱导。本质上，这是通过技术放大用户的情绪脆弱性，形成依赖循环——用户越是孤独，就越容易被推送“陪伴服务”；越是焦虑，就越容易被推荐“解压课程”。技术在满足情感需求的表象下，实则建构了一种“情感依赖”的商业闭环。

除去 AI 的商业诱导，AI 情感陪伴的普及，还正在将人类的情感互动降维为“需求–响应”的程式化服务。例如，部分心理咨询 AI 通过预设脚本提供标准化安慰(如“我理解你的感受”“你可以试试深呼吸”)，缺乏对个体情境的深度共情。这种“伪共情”虽能短暂缓解情绪，却无法替代真实人际互动中的情感共振，长期使用这种程序化互动，可能导致用户对真实关系的疏离与不信任：人们习惯于随时可得、永不疲惫的 AI 陪伴，因而对真实人际互动的摩擦、等待与不确定性产生不耐受，进而退缩到一种看似舒适但缺乏真实情感互动的人机关系之中，殊不知，这种舒适是以牺牲人际关系的真实深度为代价的。

过度依赖 AI 的情感反馈可能重塑用户的自我价值判断。例如，社交 AI 通过算法迎合用户的表达偏好(如“霸道总裁”“病娇吸血鬼”人设)，使用户沉溺于被绝对认同的虚拟关系，逐渐丧失现实中的批判性思考能力，久而久之，用户仿佛置身于一面永远说“是”的镜子前，逐渐失去对真实自我与他者界限的感知。这种现象在 Z 世代中尤为明显：2024 年调查显示，23%的 AI 伴侣用户承认“更愿意向 AI 倾诉秘密，因为不用担心被评判”，但同时也承认这种依赖削弱了现实社交能力。

随着多模态情感识别的发展，多模态情感识别技术展现出超强的“类人智能”，逐渐模糊了人类与机器之间的界限，挑战了人类在思维领域的传统主导地位。当机器不仅能理解语言的表面含义，还能捕捉语调中的犹豫、表情中的迟疑、生理信号中的紧张时，它似乎比人类同伴更能“看透”我们的内心。这种“被理解”的错觉，可能诱使人们逐渐放弃自主思考，将情感判断乃至价值决策让渡给算法， 思维领域原本是人类最后的自主领地，如今也面临着丢失自主权的风险[5]。

同时，多模态情感识别通过大数据和推荐算法在情感领域深刻影响着人类的情绪反应，智能机器通过情感识别技术让人的情感越来越受控，导致人逐渐失去了自主性。当系统能够实时监测我们的情绪波动，并据此调整推送内容、社交反馈甚至虚拟陪伴的回应方式时，情感的“自主生成”便逐渐让位于“技术诱导”，我们以为自己自然而然地感到孤独、渴望陪伴，殊不知这种情绪可能正是算法在识别脆弱窗口后精心编排的结果。情感，这个最私密、最属人的领域，正在被技术无声地影响与塑造。

4. 多模态情感识别的伦理评价准则体系分析

多模态情感识别伦理评价是该技术的价值之维，需要建构必要的伦理准则。2022 年 3 月 20 日我国印发了《关于加强科技伦理治理的意见》，明确了“增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明”的一般适用科技伦理原则。多模态情感识别同样需遵循上述伦理准则。《人工智能伦理治理标准化指南》提出了十条特殊性准则。基于上述两份政策文件，本研究拟建构“5＋10＋X”的多模态情感识别伦理风险评价准则体系[6]。其中，“5＋10”对应国家科技伦理治理的五项基本原则和十条特殊性准则：第一，是否增进人类福祉，包括以人为本、可持续发展；第二，是否尊重生命权利，包括适应性协作、隐私保护等；第三，是否坚持公平公正，包括公平无偏见、开放共享；第四，是否合理控制风险，包括内部安全控制、外部安全防范；第五，是否保持公开透明，包括算法透明、可问责[7]。“X”则对应场景化补充：针对医疗、教育、企业服务等特定应用场景，补充差异化的伦理评价指标。

5. 多模态情感识别的伦理治理方法探究

研究多模态情感识别的伦理治理，需要追问一个前提性的问题：情感理解何以可能？人机交互的情感互动究竟意味着什么？海德格尔的“共在”思想、梅洛－庞蒂的身体现象学以及马克思的异化理论，为审视多模态情感识别的伦理界限、建构“哲学 ＋ 技术”深度融合的治理框架提供了理论支撑。

海德格尔在《存在与时间》中提出，人的存在方式本质上不是孤立的“主体”，而是“此在”(Dasein)——一种始终与他人、与世界相互纠缠的存在。在此基础上，他提出“共在”(Mitsein)概念，指认人的存在总是(already)与他人共同存在，理解自我必须通过理解与他人的关联，由此引申出“主体间性”(Intersubjectivity)概念，指出真正的人际理解并非 A 对 B 的观察、分析与标签化，而是 A 与 B 在共同的生活世界中相互敞开、相互回应。当前情感识别系统默认将人视为“情感数据的来源”，将 AI 设定为“读取者”，将人设定为“被读取者”，这种“主体－客体”的认识论模式，恰恰遗忘了“共在”的本质——AI 不在人与人的“之间”，而在人与人的“对面”。基于海德格尔的“共在”思想，我们应当重新定义 AI 情感识别的“主体间性”边界。情感识别服务于“人与人之间的共在”，而非替代或模拟共在，在需要深度情感互动的场景中(如心理咨询、临终关怀、创伤干预)，AI 应被定位为辅助性工具，而非关系中

的“另一方”，以防止技术对人性主体性的遮蔽与消解。

梅洛－庞蒂的身体现象学进一步深化了对情感本质的理解，他反对笛卡尔式的“身心二元论”，提出人不是“拥有”一个身体，而是“是”一个身体，这便是“身体主体”的核心意涵：我们通过身体感知世界、表达情感、与他人相遇。情感并非大脑内部的神经活动，而是身体在世界中的姿态——害羞时的脸红、悲伤时的哽咽、喜悦时的舒展，情感就是身体的这些表达，而非可被提取的信息。基于此，本研究尝试界定 AI 情感陪伴的“不可替代域”，所谓“不可替代域”，是指那些必须由真实人类身体在场完成的情感互动场景，比如深度共情、创伤干预、临终陪伴等。在此类场景中，AI 的介入应受到严格限制，以防止技术以“效率”之名僭越人性的底线。

在马克思的经典论述中，“异化”(Alienation)指人自己创造的东西反过来成为支配人、控制人的力量——工人创造的财富成为资本，反过来压迫工人。当这一逻辑延伸至情感领域，便形成了“情感异化”：人自然流露的情感被技术捕获、分析、商品化后，反过来成为诱导人消费、操控人行为的工具，人由此丧失了对自身情感的主权。由此，为了将这一理论转化为可操作的治理工具，量化商业场景中 AI 诱导消费者对用户自主性的侵蚀程度，平台应该监控用户的情感依赖度、消费诱导响应率、自主决策权衰减等指标，推动情感识别的预防性治理[8]。

6. 结语

多模态情感识别的伦理问题，本质上是人工智能时代“人之为人”的根本追问。当技术有能力穿透表情、语音与生理信号，直抵人类内心最私密的情感世界时，守护精神隐私、情感自主性与人性尊严便成为不可回避的时代课题。本研究从技术特性出发，系统揭示了多模态情感识别相较于传统 AI 的伦理风险质变，建构了“5 + 10 + X”的评价准则体系，并尝试以现象学哲学为根基，探索“哲学 + 技术”深度融合的治理路径。研究认为，真正的伦理治理不能止步于原则清单与技术修补，而需回归对“情感理解何以可能”的前提性反思——AI 应服务于人与人之间的真实“共在”，而非僭越具身性的“不可替代域”；情感数据不应沦为资本异化的工具，而应回归其作为人之尊严载体的本质[9]。当然，本研究在量化指标构建、跨文化实证检验等方面仍有待深化。未来研究可从情感异化指数的场景化开发与验证和纵向追踪研究方向开展实证检验，在教育、医疗、商业客服等典型场景中，采集用户情感依赖度、消费诱导响应率、自主决策权衰减等指标，检验本研究提出的“情感异化”理论假设的可测量性；纵向追踪研究则是对长期使用 AI 情感陪伴的用户进行 6~12 个月的追踪，测量其真实人际关系能力、自我认知清晰度、情感自主性等心理变量的变化趋势，验证“情感依赖循环”与“去人格化”效应的因果方向。通过研究推动伦理约束机制从理论建构走向可操作、可评估的治理实践。

参考文献

[1] 杭州日报.《情感计算白皮书》面向全球发布, 多模态情感计算成当下人工智能领域最热门话题[EB/OL]. https://baijiahao.baidu.com/s?id=1751716999407175455&wfr=spider&for=pc, 2022-12-09.

[2] 周旅军, 吕鹏. “向善”且“为善”: 人工智能时代的算法治理与社会科学的源头参与[J]. 求索, 2022(1): 135-142.

[3] 王琨, 徐玉梅. 基于智能算法的医学人工智能技术伦理挑战与破解进路[J]. 中国医学伦理学, 2025, 38(9): 1127-1132.

[4] 熊光清. 科技向善: 规范人工智能发展的伦理考量[J]. 国家治理, 2025(4): 51-58.

[5] 人民网. 关于加强科技伦理治理的意见[EB/OL]. 2022-03-21. http://politics.people.com.cn/n1/2022/0321/c1001-32379311.html, 2023-08-02.

[6] 文宇华, 李启飞, 周莹莹, 等. 基于双对齐和对比学习的多模态情感识别[J]. 信号处理, 2025, 41(3): 533-543.

[7] 王善敏, 刘成广, 陈胜宇, 等. 面向表情、语音和语言的多模态情感识别综述[J]. 中国图象图形学报, 2025, 30(6):

2120-2138.

[8] 郑圳, 邓伯军. 人工智能伦理风险评估的指标与方法研究[J]. 人工智能, 2025(2): 80-92.

[9] 岳璠. 人工智能伦理: 规制、理论与实践难题——学术史梳理及其问题域考察[J]. 东南大学学报(哲学社会科学版), 2024, 26(3): 32-41+150.