# 为 AI 立“心”：人工智能时代中国社会保障的双重使命

吴 鑫

［摘 要］ 人工智能（AI）需秉持“向善”原则，确保技术进步服务人民福祉是时代的核心命题。本文提出，中国式智慧向善的本质是一种以社会保障为基石的“价值对齐”实践，让技术逻辑与社会保障制度的目的性价值对齐。文章构建“理论－历史－实践”三维框架，阐明社会保障具备为 AI 定性的道义正当性，作为社会转型“稳定器”的历史必然性，以及通过制度刚性、资源调配与全民覆盖所彰显的不可替代性。在此基础上，文章解析了社会保障与智慧向善的内在统一性，预警“道德萎缩”风险，并最终提出：应发挥社会保障作为“文化－制度”复合体的双重功能，既通过中国社会保障适应性改革守住民生安全底线，更以其目的性价值为 AI 发展提供规范，为全球 AI 治理贡献中国智慧与方案。

［关键词］ 向善治理；社会保障；人工智能；价值对齐；文化规范

## 一、引言

随着国务院“人工智能 +”行动计划的全面铺开，社会正处于一个由技术塑造未来的转型关口。[①] 这背后，是中国数字经济的蓬勃发展，其规模已从 2005 年的 2.62 万亿元增长至 2023 年的 53.90 万亿元，占国内生产总值（GDP）比重高达 42.8%。[②] 作为新质生产力的核心引擎，人工智能（AI）正从供需两侧改变着经济与社会图景：它不仅能通过优化资源配置与加速技术创新来提升全要素生产率，也致力于创造新型服务与消费，以满足人民对美好生活的多元向往。[③] 然而，技术的效率价值并不天然等同于社会伦理的“善”。若缺乏正确的价值引导，人工智能以效率为核心的技术理性就极有可能凌驾于人文价值之上，导致发展方向的偏离。因此，如何为人工智能确立“善”的价值内核，构建一个确保技术进步最终服务于全体人民福祉的“向善治理”框架，便成为这个时代的根本性社会课题之一。

［作者简介］ 吴鑫，香港大学文学院哲学系硕士研究生。主要研究方向：伦理及社会、公共政策。

① 国务院：《关于深入实施“人工智能 +”行动的意见》，中国政府网：https://www.gov.cn/zhengce/content/202508/content_7037861.htm，2025 年 8 月 26 日。

② 中国信息通信研究院：《中国数字经济发展研究报告（2024 年）》，中国信通院官网：https://www.caict.ac.cn/kxyj/qwfb/bps/202408/P020240830315324580655.pdf，2024 年 8 月 27 日。

③ 蔡昉：《为实施“人工智能 +”创建良好制度环境》，《学习时报》，2025 年 9 月 8 日第 1 版。

本文认为，在中国式现代化的独特语境中，能够为全社会人工智能立“心”的，正是国家的社会保障体系。它不应该仅仅被视为 AI 技术的一个应用领域，而应是 AI 技术向善发展的关键社会基础。这一论断是基于对现有学术讨论的反思。

人工智能与社会保障制度已成为一个全球性的学术议题，其研究主要沿着两条路径展开：一是技术如何赋能、伦理如何治理；二是制度如何转型、价值如何协同。在技术赋能与伦理治理层面，国内外学界已形成广泛共识。一方面，学者们普遍认可 AI 在优化社会服务、实现主动化管理与自动服务方面潜力巨大。[①] 国内学者进一步将 AI 演进概括为从信息化、数字化到智能化的转型，并对构建“数字社保智能体”进行了分析设想。[②] 另一方面，对伦理风险的警示是更为深刻的国际共识，聚焦于价值对齐、算法公平性、透明度与问责制等问题。其核心直指人工智能本身的技术缺陷与治理难题。在社会保障领域，国内外学者普遍警示，由于训练数据偏差或算法设计不当而产生的“算法歧视”，以及因模型复杂性导致的 “算法黑箱”问题，容易固化甚至加剧对弱势群体的不公。[③] 并且有观点指出，数字化转型伴随着对传统就业的冲击、筹资机制的模糊化以及数据安全等制度不确定性。[④] 因此，如何建立强有力的法律监管框架，以确保算法的公平、透明与可解释性，已成为全球治理的焦点，如欧盟的《人工智能法案》（AI Act），以及我国的《人工智能治理蓝皮书》《人工智能安全治理框架》均是对此的探索。

在更深层的制度转型与价值协同层面，学者们指出，为应对数字化冲击，社会保障体系需进行系统性重构，[⑤] 并尝试将此议题提升至文明演化的哲学高度，提出了哲学社会科学与 AI 应在价值、认知、实践上实现“双向赋能”，其核心在于推动哲学社会科学的价值导向与 AI 的技术逻辑深度融合，以守住“人之为人”的底线。[⑥]

综上，现有国内外研究描绘了 AI 时代社会保障变革的宏大图景，但仍存在一个关键的研究空白。无论是强调风险治理，还是呼吁价值赋能，现有研究大多在探讨如何用价值去规范技术，而较少追问这个“价值”的具体内容和制度基础究竟是什么。虽有研究深刻指出了价值导向的重要性，但其论述更侧重于哲学社会科学作为一种学科体系的普遍性功能。现有研究尚未充分揭示，社会保障体系如何能凭借其独特的属性，超越自身范畴，为整个社会的 AI 向善治理提供根本性的价值规范。本文认为，这一问题的答案正蕴含在中国社会保障体系自身之中。它不仅是一个内嵌了“共同富裕”与“人的全面发展”等“目的性价值”的文化载体，更是一个覆盖面极广的制度实体。[⑦] 正是这种文化价值与制度权力的深度融合，使其作为“文化 – 制度”

① Benouachane Hassan, "Artificial Intelligence in Social Security: Opportunities and Challenges," *The Journal of Social Policy Studies*, 2022, 20(3); Qiang Sun, "Intelligent Social Welfare: How AI Optimizes Social Assistance, Elderly Care, and Healthcare Systems," *Digital Society & Virtual Governance*, 2025, 1(1).

② 翟绍果：《数字化转型中社会保障的结构改革与制度创新》，《社会保障评论》2025 年第 2 期。

③ Terry Carney, "Artificial Intelligence in Welfare: Striking the Vulnerability Balance?" *Monash University Law Review Advance*, 2020, 46(2).

④ 蔚海燕、朱苇琦：《人工智能风险识别与治理路径构建》，《情报理论与实践》，https://link.cnki.net/urlid/11.1762.G3.20250901.0932.002，2025 年 9 月 1 日。

⑤ 陈斌：《中国式现代化进程中的数字化转型与社会保障制度适应性改革》，《社会保障评论》2025 年第 4 期。

⑥ 刑纪红、徐进：《哲学社会科学与人工智能的双向赋能——理论逻辑、实践路径与未来图景》，《南京社会科学》2025 年第 8 期。

⑦ 郑功成：《中国式现代化与社会保障新制度文明》，《社会保障评论》2023 年第 1 期。

复合体构成了本文所提出的核心分析概念，并因此天然具备为 AI 价值对齐提供本土化、实体化、制度化规范的独特功能。

为弥补上述研究空白，并深入阐释这一核心概念，本文的具体研究问题是：第一，在人工智能时代，中国社会保障体系为何能够以及如何能够凭借其“文化 – 制度”复合体属性，承担起保障民生与引领技术向善的双重使命？第二，支撑起这一双重使命的内在作用机制与实践路径是什么？

为回答上述问题，本文的核心论点是：社会保障体系在 AI 时代应扮演双重角色，这不仅是其自身发展的需要，更是实现全社会 AI 向善治理的关键。其一，是履行“适应性安全”的功能：面对 AI 带来的社会风险，通过内部机制改革，吸收和缓冲技术冲击，守护社会稳定与民生安全底线；其二，是发挥“目的性价值引领”的功能：将其内含核心价值，确立为 AI 技术开发与应用的根本伦理准则。一个真正“向善”的智能社会，必然是一个社会保障与人工智能深度融合、协同共生的社会。

为系统论证此观点，本文采用理论建构与规范分析的研究方法，构建一个“理论 – 历史 – 实践”三维分析框架。在研究路径上，本文首先从文化哲学层面溯源“善”的内涵，并论述其与中国社会保障的“目的性价值”的渊源；其次，通过历史与制度维度的考察，论证社会保障在 AI 治理中的历史有效性与不可替代性；然后，本文系统阐释社会保障与 AI 向善治理的“价值 – 功能 – 规范”内在统一性框架；最后，在上述分析基础上，提出一个“文化 – 制度”复合体的治理思路，并从适应性改革与价值引领两个层面，为发挥社会保障的双重使命提供具体路径。

## 二、“善”的文化溯源与其现代化身

“善”的内涵与标准本质上是文化的产物，而非技术所预设。现代社会保障体系正是社会将这种文化定义的“善”付诸实践的制度体现。“善”作为人类社会共通的美好向往与价值认同，其核心内涵具有普遍性，但其具体的价值排序与实现路径则带有鲜明的文化印记。尽管对幸福、公正、和谐的向往是跨文明的，但不同文明因其独特的历史、思想与传统，对“善”的理解与诠释会存在细微的差异。西方文明强调的个体自由与东亚社会推崇的集体和谐，均是对“善”的不同诠释。这些差异并无优劣之分，而是当前世界多样社会价值体系的根源。

西方哲学传统倾向于从本质与规范的视角界定“善”，强调个体权利的神圣性，将“善”视为一种可以被理性认同和普遍遵循的客观准则。例如，康德在《道德形而上学》中提出“定言命令”，要求人的行动准则必须能成为普遍法则，将“善”建立在超越个人情感的理性与普遍性之上。[①] 而尼采虽对传统道德进行了批判，但其通过“主人道德”与“奴隶道德”的区分，认为道德判断背后是特定群体的“意志”在为自身行为正名，亦凸显了西方思想对个人意志的极致关注。[②]

① ［德］伊曼努尔·康德著，张荣、李秋零译：《道德形而上学》，中国人民大学出版社，2013 年，第 15、244 页。
② ［德］弗里德里希·尼采著，周红译：《论道德的谱系》，生活·读书·新知三联书店，1992 年，第 8 页。

相较之下，以儒家思想为代表的中国哲学倾向于将“善”视为一种在人伦关系、道德实践与生命境界中实现和提升的关系性与生成性的价值。它并非一套抽象原则，而是一种需要在互动中体悟的、在于人心的道德追求。以孟子的“性善论”举例，他认为“人皆有不忍人之心”，并提出“恻隐、羞恶、辞让、是非”之心，即“仁、义、礼、智”四端。他强调：“人之有是四端也，犹其有四体也”，这表明“善”是人与生俱来的。[①] 进一步，孟子提出了“扩而充之”的实践路径，要求个体主动地、有意识地在日常生活中培育、发展和壮大这些善的萌芽，最终在和谐的社会关系与自我完善中达成“至善”境界。

儒家伦理为这份对“善”的追求提供了精密的哲学框架，它将“善”描绘成一个从内在德性萌发、经由人伦实践、最终达致人格完善的动态过程。儒家之“善”并非悬置的抽象概念，而是人格化的、生成性的价值体系。“仁”是其内核，即孟子所言的“恻隐之心”，一种推己及人、成就彼此的道德情感；“孝”则是其起点，作为最基本的人伦情感，“孝”是指个体在家庭共同体中学习关爱和践行责任，是“为仁之本”；而“德”则是“善”的追求，指个体通过长期修身，将“善”内化为言行合一的人格品性。

由此，儒家对“善”的理解，内蕴着“关系和谐”与“人格超越”两个相辅相成的维度，它们共同构筑了中华伦理的基石。其一，“善”深植于人伦网络之中，可称为“关系之善”，强调共同体的和谐与责任。这种“善”并非凭空而来，是一个始于个体内心，而成于社会关系的过程。它发端于“恻隐之心”的内在同情，并要求人们通过“克己复礼”的修身实践，将这份同情从“老吾老”的家庭内部，推及至“人之老”的广阔社会，其最终的政治理想便是“天下为公”。在此过程中，“仁、义、礼、智、信”不仅是外在的行为准则，更是维系社群和谐的内在精神纽带。与此同时，儒家之善并未消解个体的价值，反而对其提出了更高的要求，这便是“境界之善”，即个体在参与上述过程中对自身人格的塑造与超越。它激励个体以“止于至善”为毕生追求，通过“学以成人”的持续砥砺，塑造一种超越具体功用的“君子不器”式完整人格。这种人格的最终实现，在于达到“成己”与“成物”的统一，在成就自我的同时亦成就他人与世界，体现为一种自强不息的精神力量与生命价值。[②]

下面将“善”的哲学分析引向当前 AI 向善治理的核心难题：人工智能的伦理挑战与“价值对齐”问题。“善”的定义并非是统一而标准的技术参数，而是植根于特定文化的价值共识，此情景下，AI 的“价值对齐”最重要的便是与特定文化的核心价值观对齐。麻省理工学院（MIT）著名的“道德机器”实验就以海量数据证明了这一点：在全球范围内，人们对于自动驾驶汽车在极端道德困境的情境下做出的选择，存在显著的跨文化差异。[③] 这充分说明，脱离文化语境的“AI 向善”，不仅可能导致技术应用的适应性不足，更有可能引发深层次的文明冲突。须认识到任何制度设计都必须以契合国情、获得人民文化认同为前提。[④] 若忽略这一文化根基，制

① 杨伯峻：《孟子译注》，中华书局，1988 年，第 80 页。

② 杨伯峻：《孟子译注》，中华书局，1988 年，第 16 页；杨伯峻：《论语译注》，中华书局，2012 年，第 23、172 页；胡平生、张萌：《礼记（中华经典名著全本全注全译）》，中华书局，2022 年，第 1228、1757、2018 页。

③ Edmond Awad, et al., "The Moral Machine Experiment," *Nature*, 2018, 563.

④ 郑功成：《中国式现代化与社会保障新制度文明》，《社会保障评论》2023 年第 1 期。

度效能便会受制于那些隐秘但强大的传统观念。[①] 因此，构建智慧向善治理框架，必须从理解自身文明的“善”的倾向开始。

同时，对人工智能的有效治理，无法直接建立在抽象的哲学理念之上，而必须诉诸于将这些理念转化为社会契约和刚性约束的制度实体。技术的逻辑是操作性的，算法的设计需要清晰的、可执行的规则与边界。一个纯粹的哲学概念，无论多么崇高，都难以直接转化为程序员可以编码的参数或法律可以裁决的准则。因此，“向善治理”的核心挑战，就是将抽象的文化价值，具象化为可供 AI 对齐的、具有强制性与操作性的规范标准。基于此，在现代中国，这一可作为 AI 价值对齐的规范标准的、植根于文明深处的“善”的制度性载体是什么？

本文认为，答案正是国家的社会保障体系。这一制度载体的遴选，并非任意为之。审视现代国家治理体系，法律、市场与社会保障是三大核心支柱，但它们在承载“善”的价值上有所不同。法律体系以其刚性规则划定了行为的“底线”，重在惩恶与防坏，但难以主动引导和塑造“向善”的集体追求。市场机制以效率为优先，通过竞争优化资源配置，却往往因其逐利本性而加剧不平等。唯有社会保障体系，其制度内核与中华文化之“善”的追求展现出最高度的内在契合。

在现代治理层面看，一国对“善”的文化理解与相应的价值追求，集中且系统地体现就是其社会保障体系的设计与实践。换句话说，中华文化对“善”的追求，其传统理念在现代中国并未消逝，而是在吸收马克思主义、习近平新时代中国特色社会主义思想等理念后，通过社会保障制度的设计获得了新生。它是一个社会对公民尊严、社会正义与集体责任等伦理议题的制度化回应。其覆盖范围、保障水平与运行逻辑，反映了一个社会在特定发展阶段对“何为美好生活”的集体想象与承诺。中国的社会保障体系是“植根中华大地”的产物，带有中华传统文化的基因。因此，社会保障体系正是承载文化认同、并将抽象的“善”转化为公民可主张的法定权利的核心制度化体现。

那么，中国的社会保障体系具体是如何继承并体现中华文化之“善”的呢？中国现代化社会保障体系不仅是民生安全的制度保障，也并非纯粹的经济再分配工具，更是一种以“人民性”为根基的强制性社会共享机制。其关键在于它内含了一套郑功成教授所述的、超越西方纯粹工具性的独特“目的性价值”——“共同富裕”与“人的全面发展”。[②] 这代表着中华文明所追求的集体和谐、民生为本的价值理念。

一方面，“共同富裕”的目标承诺，是对“关系之善”的宏大实践。将儒家“天下为公”的共同体思想转化为由国家主导的强制性社会共享机制，以“人人为我，我为人人”的团结互助为基石，可巩固社会的整体和谐。这种将增进全体人民福祉，特别是对社会中弱势群体的人文关怀制度安排，体现了对社会公平正义的追求。这一价值追求在国际政治哲学领域亦能找到强烈的共鸣，例如，它与罗尔斯在《正义论》中提出的“差异原则”等正义理论形成深度对话。

另一方面，“人的全面发展”的价值关怀，则继承了“境界之善”的个体超越追求，它明确社会保障的功能不止于物质托底，更在于“赋能于人”，使个体从生存焦虑中解放出来，为

① 郑功成：《社会保障学：理念、制度、实践与思辨》，商务印书馆，2020 年，第 43-56 页。

② 郑功成：《中国式现代化与社会保障新制度文明》，《社会保障评论》2023 年第 1 期。

追求更高层次的精神与德性成长创造先决条件，这与“学以成人”的理想高度契合。

因此，中国的社会保障体系本身，就是一套经过现代化转译的、植根于文明深处、体现于国家制度的“善”。其能为人工智能的治理提供价值准则，根植于双重的基础之上。一是深厚的文化与历史渊源。正如前文所述，它是中华优秀传统文化中“善”的理念，在当代的延续与创造性转化。二是坚实的国家意志与政治合法性。它的“人民性”根基，与中国共产党的根本宗旨和“以人民为中心”的发展理念内在统一。党的二十届三中全会审议通过的《中共中央关于进一步全面深化改革 推进中国式现代化的决定》将“促进社会公平正义、增进人民福祉”确立为全面深化改革的出发点和落脚点，为人工智能的“向善治理”提供了根本的价值遵循和清晰的行动指南。[①] 这与习近平总书记所强调的“让国家发展成果更多更公平惠及全体人民”和“让人民生活得更加美好”一脉相承。深植于文化、确证于实践、并充分体现国家意志的社会保障在法理与道义上，天然具备了为人工智能这一新兴技术定性、立“心”的资质与权威。

## 三、社会保障的历史有效性与不可替代性

中华文明对“善”的制度性求索源远流长，其早期形态体现为一种维系社会稳定的“仁政”实践。在古代中国，当面对天灾人祸与社会动荡对统治秩序构成威胁时，统治阶级出于维护稳定的需要，发展出救灾济贫、优抚恤孤等主张与行动。这虽是被动形成的早期社会保障政策，却客观上开启了以国家力量推行社会团结与互助的先河，并被赋予“仁政”的道德意涵。在此基础上，经过漫长历史的沉淀，逐渐衍生出“天下为公”的大同社会理想、守望相助的社会互助思想以及扶危济困的社会救济传统，这些共同构成了现代中国的社会保障理论的文化渊源。[②]

从“仁政”思想到“大同”理想，再到现代的社会保障网络，国家主导的社会共济始终是应对社会重大变革的“稳定器”，其存在具有深厚的历史与文化根基。从古代的常平仓措施到现代的社会保障体系，历史反复证明，在重大社会变革时，无论是天灾人祸还是技术革命，依赖强有力的社会保障体系进行风险缓冲与秩序重构，是实现稳定转型的成功经验与惯性选择。因此，通过国家主导的社会保障来应对风险、维系团结，是中国社会治理一以贯之的核心逻辑。这也为今日应对人工智能这一颠覆性技术变革，提供了历史镜鉴和路径遵循。

在追求 AI 与文化之善对齐的进程中，社会保障体系之所以具有核心地位，是因为相较于其他社会系统，社会保障在履行 AI 治理的“指导性”作用时具备三大不可替代的独特性。

第一，其“制度刚性”为抽象的“善”提供了强制性、可操作的现实规范。文化价值往往是柔性的、倡导性的。而社会保障通过法律法规、财政税收、公共服务等刚性制度，将“善”转化为社会成员可主张的法定权利和社会必须承担的法定义务。这种将伦理理想转化为制度约束的特性，为 AI 的价值对齐提供了清晰无误的规范性目标。

第二，其“资源调配”功能是实现“善”的物质基础。社会保障是社会最主要的资源再分

① 中共中央：《关于进一步全面深化改革 推进中国式现代化的决定》，中国政府网：https://www.gov.cn/zhengce/202407/content_6963770.htm，2024 年 7 月 21 日。

② 郑功成：《社会保障学：理念、制度、实践与思辨》，商务印书馆，2020 年，第 43-56 页。

配机制之一。它通过直接的现金转移、服务提供和机会创造，将关于“公平”和“共享”的抽象价值转化为具体实践。当 AI 与这一核心的社会资源分配逻辑相结合时，便能确保技术红利转化为人民可感的福祉。这一转化过程建立在我国数十年信息化建设的坚实基础之上。以“金保工程”的实施为标志，我国初步构建了全国统一的社会保险信息系统，为今天更高阶的数字化与智能化转型奠定了数据与流程基础。[①] 近年来，我国已高度重视利用数字技术提升社会保障治理能力，相关实践已从顶层设计走向落地应用。在宏观层面，《“十四五”国家信息化规划》明确提出推进居民服务“一网通办”，利用大数据、人工智能等技术实现精准识别与服务，并以社会保障卡为载体，推动居民服务“一卡通”在政务服务、社会保障、城市服务等领域的线上线下应用。在实践层面，智能化应用已初见成效。例如，浙江省借助人工智能创建的“工伤鉴定智治”新模式，大大提升了劳动能力鉴定的效率与便捷性；[②] 广西壮族自治区构建的“156”就业公共服务体系，则依托数字平台，致力于实现对重点人群的精准就业帮扶。[③] 此外，“社保电子地图”与大数据平台的建设，也使得对特定保障对象的精准识别与服务推送成为可能，让社会保障的资源配置更加优化。[④] 这些探索共同印证了社会保障系统作为承载技术应用的制度平台，是实现技术效率向民生福祉转化的关键渠道。

第三，其“全民覆盖”属性使其成为“社会共识”的最终体现。社会保障关乎所有公民的基本利益，其制度设计是各种社会力量、价值观念经过长期博弈后形成的“共识”。因此，它所承载的“目的性价值”，是已经获得法律确认、具有广泛民意基础的、正在实践中的社会契约。这一属性在数字化时代得到了前所未有的强化。根据人力资源社会保障部新闻发布会公布的数据，截至 2024 年底，中国社会保障卡持卡人数达到 13.89 亿人，覆盖 98% 以上人口，其中 10.7 亿人同时在手机中领用电子社保卡，普遍实现居民服务“一卡通”应用。[⑤] 没有任何一个商业平台或社会系统，能够拥有如此广泛且深入的覆盖能力。这让社会保障体系能够成为确保 AI 技术红利实现全民普惠的、最权威和最可靠的制度载体。因此，以这份“社会契约”来校准 AI，其合法性、权威性和可行性远非其他原则可比。

正是这三大特性——制度刚性、资源属性和全民覆盖性，共同构成了社会保障作为一种独特的“文化 – 制度”复合体的核心内涵，它不仅是文化价值的反映，更是承载文化价值的制度支柱、物质载体和共识基础，使其在 AI 向善治理中具备了不可或缺的指导性地位。

## 四、社会保障与智慧向善的内在统一性

人工智能的“向善治理”与社会保障体系并非两条互不相干的平行线，而是共享同一逻辑

① 翟绍果：《数字化转型中社会保障的结构改革与制度创新》，《社会保障评论》2025 年第 2 期。

② 《浙江借力人工智能创建“工伤鉴定智治”新模式》，人力资源和社会保障部官网：https://www.mohrss.gov.cn/SYrlzyhshbzb/shehuibaozhang/gzdt/202307/t20230726_503602.html，2023 年 7 月 26 日。

③ 《广西：“156”就业公共服务体系助力高质量充分就业》，人力资源和社会保障部官网：https://www.mohrss.gov.cn/SYrlzyhshbzb/dongtaixinwen/dfdt/202507/t20250718_549694.html，2025 年 7 月 18 日。

④ 中央网络安全和信息化委员会：《“十四五”国家信息化规划》，中国政府网：https://www.gov.cn/xinwen/2021-12/28/5664873/files/1760823a103e4d75ac681564fe481af4.pdf，2021 年 12 月 28 日。

⑤ 翟绍果：《数字化转型中社会保障的结构改革与制度创新》，《社会保障评论》2025 年第 2 期。

内核的有机共同体。其内在统一性，可以通过一个从抽象到具体的“价值 – 功能 – 规范”三维框架得到展现。这个框架揭示了二者从“为何协同”的哲学原点，到“如何协同”的实践机制，再到“以何为准”的伦理标尺的逻辑链条。这种统一性同时呼应了中国特色社会主义治理逻辑的内核。在中国式现代化的进程中，坚持党的全面领导与坚持以人民为中心是内在统一的。[①] 同样，在人工智能这一具体领域，社会保障体系与智慧向善的协同，即国家意志与技术伦理协同，正是这一宏大治理逻辑的具象化体现。

首先，价值层面的统一。内在统一性的逻辑起点，在于二者共享同一个最终的价值目标——促进“人的全面发展”与“共同富裕”，这是其内在统一性的哲学基础。社会保障体系是对文化之“善”的制度化表达，而智慧向善则是对文化之“善”的技术化赋能。尽管路径不同，但它们共同指向一切发展最终须回归“人”本身这一价值原点。中国的社会保障体系，源于儒家的“仁政”与“民本”思想，其现代化表达是促进“社会公平正义”与“人的全面发展”，它不仅要防止人因风险而陷入困境，更要赋能于人，助其追求更有尊严与丰富的生活。同样，AI 的“向善治理”，其最高阶的要求绝不仅仅是让技术“安全无害”，而是引导技术服务于人民的整体利益与个人的自由发展。因此，在价值层面二者是一致的。

其次，功能层面的互补。如果说价值层面的统一确立了二者协同的“为什么”，那么功能层面的互补则具体回答了“如何做”。这种互补形成了一个“风险缓冲 – 效率提升”的良性循环，这是其内在统一性的实践体现。一方面，社会保障体系的基础性保障功能为 AI 的广泛应用与社会整合提供了稳定机制。人工智能革命极易带来的结构性失业与社会不平等冲击，需要社会保障通过失业救济、再就业培训、社会救助等机制来缓冲与化解。若没有社会保障这一坚实的社会安全网，任何关于技术“向善”的宏大叙事都可能因社会动荡而失去意义。另一方面，人工智能的“赋能”功能则为社会保障体系的现代化提供了强大技术引擎，使其能更高效与精准地实现“善”的目标。国内外经验均已证明，AI 技术在社会服务和福利管理的精确化、自动化、主动化等方面具有巨大优势。[②] 如我国在工伤鉴定、就业帮扶、基金监管等领域的智能化实践，已初步印证了这种技术赋能正从蓝图走向现实。社会保障为 AI 的创新兜底，AI 为社会保障的价值实现增效，二者一体两面。

最后，规范层面的校准。从共享的价值目标，到互补的功能机制，最终必然导向一个关键问题：在协同过程中，以谁的标准为准？这就进入了内在统一性的最高层次——规范层面的校准。技术本质上是工具性的，其自身并无固定的价值方向。而任何技术设计逻辑都具有“非中立性”，必然渗透着特定的价值判断。[③] 因此，AI 的“向善”并非一个自然而然的过程，它必须被规范在一个坚实且明确的价值体系之上。而中国社会保障体系所承载的、内嵌了民族文化基因的“目的性价值”，正能为技术逻辑提供经过社会契约确认的、具有合法性的价值规范，

① 杜飞进：《深刻把握坚持党的全面领导与坚持以人民为中心的内在统一性》，《光明日报》，2025 年 3 月 3 日第 6 版。

② Benouachane Hassan, "Artificial Intelligence in Social Security: Opportunities and Challenges," *The Journal of Social Policy Studies*, 2022, 20(3); Qiang Sun, "Intelligent Social Welfare: How AI Optimizes Social Assistance, Elderly Care, and Healthcare Systems," *Digital Society & Virtual Governance*, 2025, 1(1).

③ 刑纪红、徐进：《哲学社会科学与人工智能的双向赋能——理论逻辑、实践路径与未来图景》，《南京社会科学》2025 年第 8 期。

确保 AI 的发展始终朝向提升人的福祉与巩固社会团结的正确方向。

## 五、“道德萎缩”与协调共生的必要性

在确立了社会保障与智慧向善的内在统一性后，更进一步的是这一内在统一性的必要性。本文认为，人工智能以效率和计算为核心的工具理性，在缺乏正确价值引导时，恐将引发一种系统性的“道德萎缩”，对中华文化之“善”构成侵蚀。

这种“道德萎缩”，首要体现在对“关系之善”的实践性侵蚀。当技术对人际责任过度替代，个体履行道德责任的主动性、情感投入乃至实践能力都可能出现衰退。其核心在于伦理实践的职能转移。以养老为例，一个功能强大的 AI 养老机器人，或许能高效完成照护工作，但这种高效的“算法处理”，与真正的情感关怀之间存在着本质的鸿沟。正如约翰·塞尔在其著名的“中文房间”思想实验中所揭示的，能够完美地遵循规则处理信息，绝不等于拥有真正的理解与共情。[①] 当子女将本应投入感情与责任的伦理互动委托给一个没有理解能力的系统时，伦理关系本身也就被悬置，最终变得脆弱。

这种对伦理实践的侵蚀，不止作用于社会关系层面，更会危及“境界之善”的实现，催生“人的平庸化”风险。在一个由算法主导的高度技术化、便利化的环境中，个体过度依赖系统，容易逐渐丧失内在的成长动力、批判性思维以及追求“止于至善”的卓越志向。AI 的高度便利性易造成“理性化平庸”，容易使人规避挑战、疏于学习，这与儒家“君子不器”的人格理想背道而驰。若 AI 赋能的社会，最终倾向于将人塑造成更高效的“数据单元”，而非更完善的“人”，那么“人的全面发展”这一终极目标将被悬置，个体的精神世界亦将趋于扁平。这不仅违背了社会保障理论中“以人的自由全面发展作为价值追求”的根本目标，也挑战了在技术时代“坚守人之为人的底线”这一呼吁。[②]

然而，将风险简单归咎于技术本身，便会陷入技术决定论的误区。AI 并非必然导致伦理空洞化。相反，若能将其定位于承担重复性劳动，反而能将人从疲惫中解放，为更高质量的情感互动创造空间。AI 的最终影响，取决于其社会应用场景与制度设计，真正的风险并非源于技术本身，而是一种将技术视为“替代”而非“赋能”的思维模式，包括对社会关系培育的忽视以及对技术应用的过度简化。有学者在反思澳大利亚的 AI 实践时指出：许多 AI 系统的失败源于其设计中对人类复杂性与脆弱性的理解不足。[③] 当前许多数字化转型仅仅是将线下业务“复制粘贴”到线上，并未触及根本的思维模式与价值逻辑。[④] 这种将复杂的社会互动简化为技术流程的思维，正是“道德萎缩”的温床。一旦维系“共同富裕”的社会情感联结与团结互助精神因实践的架空而被侵蚀，社会保障所追求的团结与和谐便会被动摇根基。这便反证了，将人工

① John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, 1980, 3(3).

② 郑功成：《在法治轨道上推进数字治理》，《学习时报》，2023 年 6 月 14 日第 1 版。

③ Terry Carney, "Artificial Intelligence in Welfare: Striking the Vulnerability Balance?" *Monash University Law Review Advance*, 2020, 46(2).

④ 杨立雄：《数字化转型与“创造性破坏”：社会保障数字治理研究》，《社会保障评论》2023 年第 5 期。

智能的技术发展置于社会保障的价值框架内进行协调与引导，是何等必要。

同时，须避免价值决定论的倾向。本文基于儒家伦理构建的“向善”框架，旨在为 AI 治理提供深厚的文化根基，而非一套僵化封闭的答案。在强调智慧向善时，必须警惕任何价值叙述的“绝对化”风险，智慧治理需防止“集体之善”对个人空间的过度挤压，尊重个体选择的多样性。例如，理解“躺平”等现象背后可能蕴含的对过劳文化的消极抵抗和生活本真的朴素追寻。向善治理的真正目标绝非塑造千人一面的“君子”，而是依托社会保障，为多元、有尊严的人生选择提供支撑。在集体和谐与个体自由之间，在激励奋进与保障尊严之间，寻求一种动态的平衡。

## 六、治理之道：“文化 – 制度”复合体在 AI 治理机制中的双重路径

“智慧向善”的治理之道，关键在于充分发挥中国社会保障体系作为“文化 – 制度”复合体的双重功能。社会保障体系并非被动地等待被 AI 改造，而是能够主动地成为 AI 向善治理机制的核心组成部分。这主要通过两条相辅相成的路径得以实现。一是履行“适应性安全”的功能，以“底线防御”逻辑构建风险缓冲机制；二是发挥“目的性价值引领”的功能，以“高线引领”逻辑为技术发展设定伦理规范。

其一，底线防御路径是指社会保障作为社会风险抵御机制的适应性改革。坚实的制度根基是实现价值引领的前提。面对人工智能浪潮带来的系统性社会风险，社会保障体系的首要任务是进行前瞻性的适应性改革，扮演好社会安全网的角色。首先，改革必须直面 AI 对就业结构与收入分配的巨大冲击。数字化转型正从根本上冲击着传统就业结构，并因其平台化的特性而模糊了社会保障的筹资机制与责任主体。[①] 例如，据新华网报道，微软在 2025 年 5 月与 7 月的极短时间内裁员了超 1.5 万人，类似的裁员趋势也出现在其他科技巨头中，如 Meta 和 Amazon 等。世界经济论坛《2023 年未来就业报告》的预测更为严峻，指出到 2027 年全球或将净减少 1400 万个工作岗位，行政与秘书类岗位是受冲击最严重的领域之一。这种风险不仅体现在宏观层面，更体现在对个体尤其是弱势群体的排斥上。设计不当的 AI 系统极易对技术匮乏或处境复杂的弱势公民造成事实上的歧视，放大而非弥合社会鸿沟。[②] 因此，社会保障体系的适应性改革，首要任务就是守住公平、安全底线，确保技术进步不会以牺牲任何一个群体的福祉为代价。为此，社会保障的适应性改革必须多管齐下。一方面，要利用好 AI 技术赋能社会保障的高效化、精细化管理。比如，利用 AI 技术加固社会保障基金安全的“防火墙”：基于大数据技术建立常态化的疑点数据筛查机制，利用 AI 风控规则实现对高风险业务的毫秒级响应与智能监控，确保安全网自身的可持续性。宁夏等地推行的“数字安全员”制度，正是将智能化监管嵌入经办

① 陈斌：《中国式现代化进程中的数字化转型与社会保障制度适应性改革》，《社会保障评论》2025 年第 4 期。

② Terry Carney, "Artificial Intelligence in Welfare: Striking the Vulnerability Balance?" *Monash University Law Review Advance*, 2020, 46(2).

流程的生动实践，有效降低了基金的损失风险。[①] 另一方面，更须探索与数字经济相适应的新型制度安排。例如，国际上探索的数字社会保障（DSS）模式，即依托数字平台直接为平台工作者组织社保，为破解新业态劳动者参保难题提供了有益启示。[②] 可借鉴这一思路，在此基础上探索适合我国国情的平台化缴费机制。此外，改革还要聚焦于“赋能于人”与“成果共享”。尽管 AI 会创造新的岗位，但劳动力市场将面临巨大的结构性转变和技能不匹配问题。可由社会保障机构牵头构建一个覆盖全民的终身学习与再培训体系，赋能劳动者适应技术变革，以应对结构性失业。尽管构建系统性的终身学习体系任重道远，但在特定时期，我国已展现出利用数字平台整合培训资源的行动力与制度优势。例如，在新冠疫情防控工作期间，为实施“互联网 + 职业技能培训计划”，人力资源和社会保障部便推荐了 54 家职业技能培训线上平台机构，免费提供线上职业技能培训资源及服务。最后，应探索建立更有效的数字红利分享机制。例如基于数字产权的数据税、机器人职工的自动化税等新型的筹资模式。中国共产党所具备的“总揽全局、协调各方”能力，为推进数字红利的公平共享提供了坚实的政治保障。[③] 要把党的二十届三中全会《决定》中“改革成果由人民共享”的原则落到实处，确保发展的成果得到更充分、更公平的分享。

其二，高线引领路径是指发挥社会保障“目的性价值”的规范性作用。坚实的制度支撑仅是起点，其最终目标是指向由“目的性价值”所提供的文化规范，以引领人工智能发展的长远价值航向。社会保障不应满足于基础性保障功能，更要主动成为一个前瞻性的规范性框架。在国际上，一种主流的治理思路是以欧盟《人工智能法案》（AI Act）为代表的基于风险的底线规制路径。该法案将 AI 在社会保障等领域的应用划定为“高风险系统”进行严格监管，同时，辅以《数字市场法案》（DMA）等法规强制要求平台算法透明化。[④] 我国也先后出台了《人工智能治理蓝皮书》《人工智能安全治理框架》等治理措施。这种外部监管模式重在划定“不可为”的法治红线。而本文提出的以社会保障为价值内核的中国特色方案，则倡导一种“内生引领”模式。在确立底线的基础上，更要通过将“共同富裕”和“人的全面发展”等积极价值嵌入技术的设计与应用全过程，树立一个“应当为”的价值高线。这是一种从被动风险防范到主动价值塑造的治理模式升级。这种引领并非空谈，在社会救助这一最能体现社会温度的领域，智能化实践已在探索如何更好地实现“善”的目标。例如，一些地方探索的“救助通”等数字化社会救助平台，正通过简化申请流程、精准识别需求、保护求助者隐私，使社会救助变得更高效和更精准、救助对象更有尊严。[⑤] 其评价标准的核心便是人民的“获得感”。正如习近平总书

① 《宁夏推进社保经办管理数字化转型》，中国政府网：http://www.gov.cn/flanbo/sffang/202502/content_7003357.htm，2025 年 2 月 11 日。

② Enzo Weber, "Digitale Soziale Sicherung: Potenzial für die Plattformarbeit," *Wirtschaftsdienst*, 2020, 100(1).

③ 杜飞进：《深刻把握坚持党的全面领导与坚持以人民为中心的内在统一性》，《光明日报》，2025 年 3 月 3 日第 6 版。

④ Lena Enqvist, "Rule-based versus AI-driven Benefits Allocation: GDPR and AIA Legal Implications and Challenges for Automation in Public Social Security Administration," *Information & Communications Technology Law*, 2024, 33(2).

⑤ 杨立雄、刘曦言：《从“情景”到“脱域”：数字时代的福利身份与社会权利探讨——基于 C 市“救助通”的案例研究》，《学术研究》2024 年第 9 期。

记所强调的，要把“是否给人民群众带来实实在在的获得感”作为改革成效的根本评价标准。[①] 将“善”的标准从哲学思辨带入技术研发与应用的一线，可确保技术进步最终服务于人的尊严与精神的丰富。

综上，社会保障体系的适应性改革与价值引领，共同构成了中国式“智慧向善”治理的一体两面。前者解决了制度韧性的现实问题，后者则回答了路径选择的方向问题。这一根植于自身文化传统与制度优势的治理框架，并非被动地适应技术，而是主动地以文明价值去塑造技术，为全球人工智能治理贡献出充满东方智慧与文化底蕴的中国方案。

## 七、总结与展望

本文研究表明，实现人工智能的“向善治理”，关键在于推动其与作为中华“善”文化现代化制度载体的中国社会保障体系的深度融合。中国的社会保障体系并非单纯的民生安全网，而是一个内嵌“共同富裕”与“人的全面发展”等“目的性价值”的“文化 – 制度”复合体。它在人工智能时代扮演着至关重要的双重角色：既要通过前瞻性的适应性改革，为技术变革的社会提供缓冲与保护；又要以其深厚的价值内涵，为智能科技划定伦理轨道，树立价值标杆。

本文的主要理论贡献包括如下方面。第一，提出了社会保障作为“文化 – 制度”复合体的分析视角。该视角揭示了中国社会保障体系不仅是风险管理工具，更是承载中华民族“善”文化现代化表达的制度载体，其内含的“目的性价值”能够为 AI 的价值对齐提供本土化的实体规范，也理应成为 AI 设计与应用的评估指标。第二，本文系统阐释了该复合体在人工智能时代应履行的“适应性安全保障”与“目的性价值引领”的双重使命，并构建了诠释二者内在联系的“价值 – 功能 – 规范”三维统一性框架，为理解社会保障体系在 AI 向善治理中的角色与功能提供了一种理论解释。

基于上述研究，本文提出如下政策建议。在筹资机制上，可以启动对“自动化税”“数字服务税”的可行性研究，探索建立与智能经济形态相适应的社会保障筹资新模式，确保技术红利能够反哺社会安全网；在再培训上，建议由社会保障部门牵头，整合教育、企业等领域资源，建立覆盖全民的“技能提升教育服务”体系，为劳动者适应 AI 带来的职业转型提供制度化支持；在数字红利分配上，探索建立国家主导的公共数据基金，将部分由人工智能应用产生的数据价值收益，以补充医疗保险、养老金等形式，实现全民共享；在治理法律法规层面，应在强化人工智能法治时，明确将“增进社会福祉、促进社会公平”等源自社会保障的核心价值原则作为 AI 开发与应用的基本准则，并建立算法的社会影响评估制度。为有效落实这些思路，应鼓励利用 AI 监管沙盒，借助数字孪生技术对相关政策或应用进行前瞻性模拟，同时设计“AI 伦理清单”，引导技术提供方进行伦理自查，并将群体误判率、舆情敏感度等伦理风险指标纳入沙盒监管的评估体系中。

① 杜飞进：《深刻把握坚持党的全面领导与坚持以人民为中心的内在统一性》，《光明日报》，2025 年 3 月 3 日第 6 版。

# A Moral Compass for AI: The Dual Mission of China's Social Security System in the Age of Artificial Intelligence

Wu Xin

(Department of Philosophy, Faculty of Arts, The University of Hong Kong, Hong Kong 999077, China)

**Abstract:** The governance of Artificial Intelligence (AI) should adhere to the principle of "doing good", making it a central task of our time to ensure that technological progress serves public well-being. This paper argues that the essence of a Chinese-style vision of "intelligence for good" is a practice of "value alignment" grounded in the social security system, aligning the logic of technology with the system's embedded purposive values. To substantiate this thesis, the paper constructs a three-dimensional "theory–history–practice" analytical framework. It elucidates the system's moral legitimacy to define values for AI, its historical necessity as a "stabiliser" in social transformations, and its practical irreplaceability, demonstrated through its institutional robustness, resource allocation capabilities, and universal coverage. Building on this, the paper analyses the intrinsic connection between social security and "intelligence for good," warns of the risk of "moral atrophy" if the two become decoupled, and ultimately proposes that the social security system should leverage its dual function as a "cultural–institutional complex": on the one hand, through adaptive reforms to safeguard the bottom line of livelihood security, and on the other, by using its purposive values to provide cultural norms for the development of AI, thereby contributing Chinese wisdom and proposals to global AI governance.

**Keywords:** ethical governance; social security; artificial intelligence (AI); value alignment; cultural norms

（责任编辑：郭　林）