CCID 赛迪智库 译丛

2019 年 4 月 29 日　　第 16 期　　总第 399 期

可信赖的人工智能道德准则草案

【译者按】2018 年 12 月，欧盟委员会公布了《可信赖的人工智能道德准则草案》。草案阐述了人工智能应遵循的道德目标，包括基本权利、原则与价值观；分析了实现可信赖的人工智能的要求；并制定了具体的评估标准。草案为开发、部署和使用人工智能的政府、社会组织、研究机构、企业和个人提供了一个人工智能道德准则框架，以便实现“以人为本”和可信赖的人工智能技术，并创造一种“可信赖的人工智能源自欧洲”的文化，使欧洲成为引领全球人工智能的创新者。赛迪智库集成电路研究所对该草案进行了编译，期望对我国有关部门有所帮助。

【关键词】欧盟　人工智能　道德准则　草案

一、概述

本文是欧盟委员会人工智能高级专家组（AI HLEG）撰写的人工智能道德准则草案。

人工智能（AI）是当代最强大的变革力量之一，势必会改变人们的社会结构。人工智能将带来繁荣发展的巨大机遇，欧洲绝不能等闲视之。得益于可用数据的海量骤增、更强大的计算结构以及机器学习等人工智能技术的突飞猛进，人工智能在过去 10 年里取得重大进步。在人工智能的推动下，自动驾驶汽车、医疗、家用/服务型机器人、教育、网络安全等领域实现重大发展，人类生活水平也随之不断提高。此外，人工智能还是解决全球重大挑战的关键手段，如全球健康与福利、气候变化、可靠的法律与民主系统以及载入《联合国可持续发展目标》的其他挑战。

人工智能拥有造福人类与社会的卓越能力，同时也会带来风险，必须妥善管理。但总体而言，人工智能仍然利大于弊，而本文的任务就是确保扬其所长、避其所短，发挥人工智能的最大优势。为确保人工智能的发展不偏离轨道，就需要坚持以人为本，必须始终牢记人工智能的发展与应用本身并不是目的，其目的是造福人类。实现可信赖的人工智能是本准则的根本宗旨，因为对于人工智能来说，人类只有信其所能，才能用其所及。

可信赖的人工智能有两个必备要素：必须尊重基本权利、适

用规则以及核心原则和价值观，确保“目的合乎道德”；在技术上必须稳定可靠，因为技术如果不受控制，好心也可能办坏事。

本准则的对象是开发、部署或使用人工智能的所有相关方，包括企业、组织、研究人员、公共服务机构、大学等社会机构、个体以及其他实体。

最后，这些指导原则不仅仅着眼于欧洲，还希望能够在全球引发人工智能道德框架的思考与讨论。

二、可信赖的人工智能道德准则框架

这份人工智能道德准则草案共分三部分，每部分均提供了指导原则，共同形成实现可信赖的人工智能道德准则框架:

道德目标

本部分侧重于人工智能各相关方所必须遵守的核心价值观与原则，其基础是在《欧盟条约》以及《欧盟基本权利宪章》规定的价值观和权利中奉为真理的国际人权法律。本节与其他章节共同构成了人工智能开发者、部署者和使用者的“道德目的”，其中应当包括尊重这些内容中规定的权利、原则和价值观。

实现可信赖的人工智能

在第一部分道德目标的基础上，第二部分提出了非穷尽的初步评估内容和要求，供人工智能的开发者、部署者和使用者从实

际操作上实现可信赖的人工智能。由于人工智能应用的特殊性，这个评估内容还需要针对具体应用、背景和领域进行修改。

可信赖的人工智能评估标准

第三部分给出了落实这些要求的可操作细则，为可信赖的人工智能提出了具体评估标准，可根据具体情况调整使用。

本准则的框架请参见图 1。

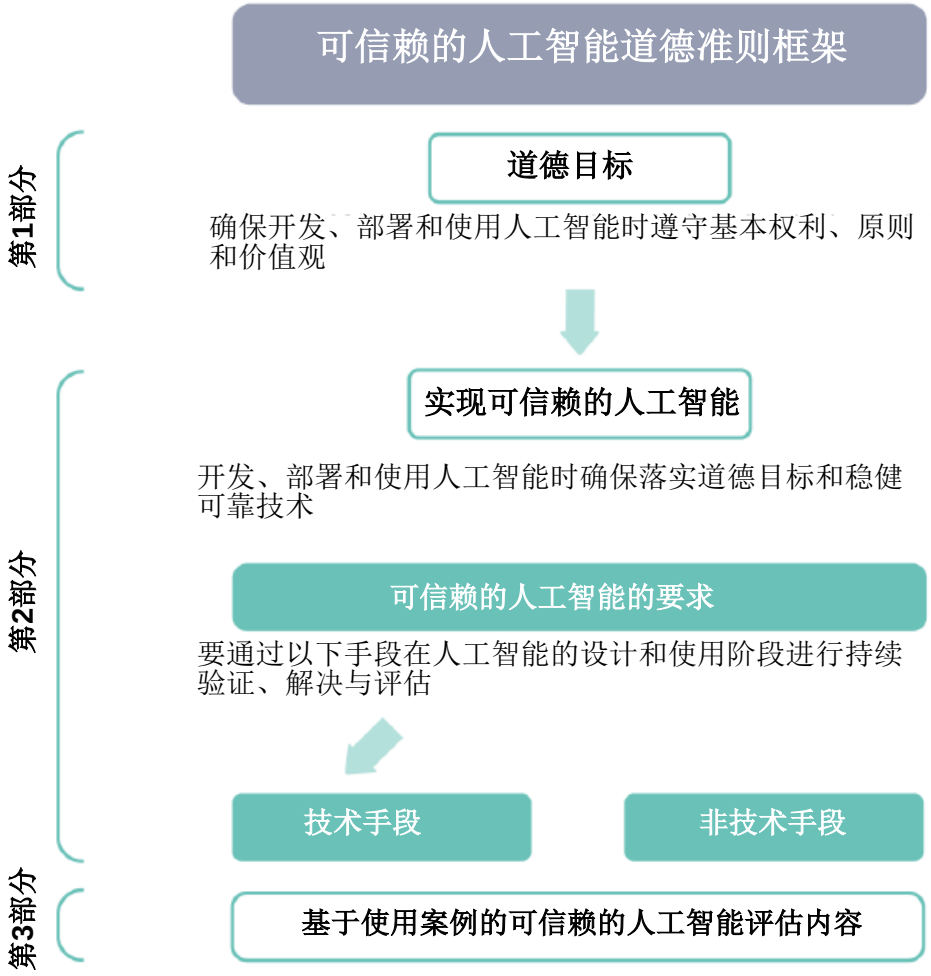

图 1：可信赖的人工智能道德准则框架

（一）道德目标：尊重基本权利、原则和价值观

1、欧盟实现人工智能道德的权利型策略

人工智能高级专家组认为，以《欧盟条约》和《基本权利宪章》中规定的基本权利为基石，找出抽象道德原则，确定如何在人工智能的语境中落实道德价值，这才是落实人工智能道德的正道。欧盟的基本义务，就是通过宪政保护不可分割的基本人权，确保尊重法律规则，推动民主自由，促进共同发展。基本权利不仅能够催生新的法律与监管规则，也是指导人工智能系统开发、使用与实施的依据。

欧盟条约和宪章规定了落实欧盟法律时所适用的权利，即：尊严、自由、平等、团结、公民权利与正义。这些权利在宪章的后续章节中进行了阐述，而贯穿其始终的主线，就是欧盟坚持以人为本的精神，即公民在政治、经济与社会工作中始终占据特殊的首要位置。

道德领域也是为了保护个体权利和自由，同时造福社会与公益。道德视角能够帮助人们理解技术如何在人工智能的开发与应用中将基本权利纳入考虑因素，以及如何更好的指导人们利用技术造福公益，而不是（现在）只关注技术能够做什么。在人工智能领域坚持基本权利，重视需要遵循的道德原则。按照这个思路，道德既是人类公认基本权利的基础，也是其补充。

人工智能高级专家组认为，以权利为基础的人工智能道德策略还能减少监管上的不确定性。欧盟几十年来统一践行基本权利的经验为使用者、投资者和创新者提供了清晰、易懂、具有前瞻性的参考。

2、从基本权利到原则和价值观

要举例说明基本权利、原则与价值观的关系，就让人们来思考一下“尊重人类尊严”的基本权利概念。这种权利涉及到对人类固有价值的认识(例如人的价值与生俱来，不关乎外貌、工作、生活或者所在的国家），如此又可引出自治的道德原则，即人人都有自由选择身体、情感或精神等方面生活方式的权利（也就是说，人的价值既然是与生俱来，那么人人都有权选择自己的生活）。继而，知情同意就是自治原则在实践中的具体操作。知情同意要求个体在决定是否开发、使用或投资实验或商用人工智能系统时，要为其提供充分的信息（如，确保人们有机会决定是否接受产品或服务，保障人们的选择权和价值观）。

这种关系虽然具有线性的表象，但实际上价值观往往要优于基本权利和/或原则。

简而言之，基本权利是形成道德准则的基石。这些原则是高级的抽象标准，开发者、部署者、使用者和监管者要奉行以人为本和可信赖的人工智能的目标，就必须遵守。而价值观是奉行道

德原则的具体指导，同时也为基本权利提供支撑。这三者之间的关系可见图 2。

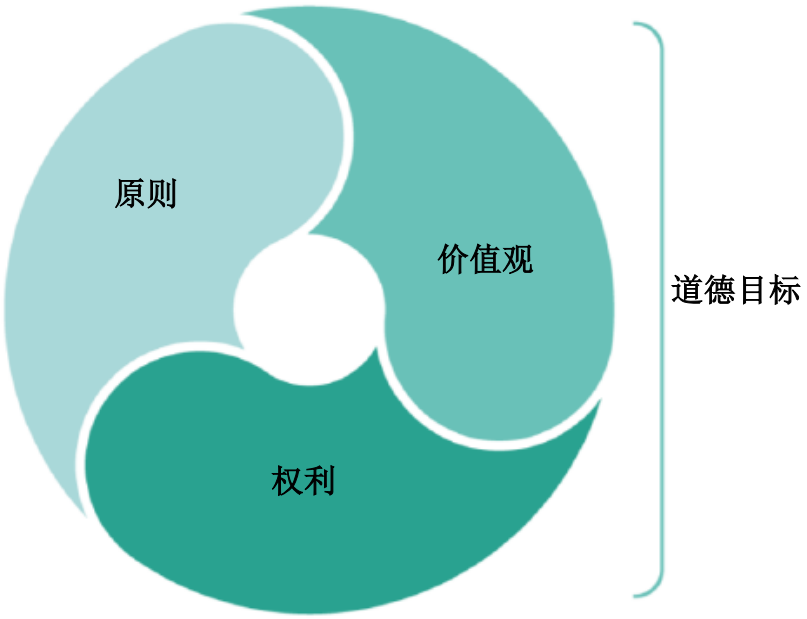

图 2：道德目标的构成要素：权利、原则与价值观之间的关系

运用基本权利推导出道德原则和价值观的做法，并非人工智能高级专家组的首创。早在 1997 年，欧洲理事会就出台了名为《在生物学和医学应用领域保护人权和人类尊严的公约》（简称奥维耶多公约）。奥约耶多公约明确规定，基本权利是在技术变革中确保“以人为本”的根本所在。

尊重并确保人工智能系统遵守基本权利、原则与价值观，就是本文中所述的确保“道德目标”，这是实现可信赖的人工智能的关键因素。

4、人工智能领域的道德原则及相关价值观

许多公共、私人与民间组织已经在运用基本权利来构建人工智能的道德框架。在欧盟，欧洲科学与新技术伦理工作组（EGE）以欧盟条约和欧盟基本权利宪章中规定的基本价值观为基础，提出了九大基本原则。近期，人工智能为人（AI4People）研究项目[1]对欧洲科学与新技术伦理工作组提出的九大原则以及迄今为止提出的其他三十六个原则进行了评估，并将其归纳为四大总体原则，包括：善行、无伤害、个体自主和公正平等。为适应人工智能的发展，专家组还对上述四大原则进行了修正，增加了第五项原则，即：可解读原则。人工智能高级专家组相信会聚的好处，这样能够找出迄今为止各种机构提出的大部分原则，同时澄清这些原则所针对的目的。最重要的，这些总体原则能够为核心价值观提供了具有可操作性的指导原则。

同时还应当注意，在特殊情况下，个体与社会对这些原则的态度可能会出现冲突。这种情况并没有特定的解决途径。在上述情况下，参照欧盟条约和宪章所保护的原则与总体价值观和权利不无裨益。鉴于人工智能的未知潜力与意外后果，本文建议要有内部与外部（道德）专家参与人工智能的设计、开发与应用。这些专家有助

[1] AI4People 项目于 2018 年 2 月由欧洲议会批准启动，旨在把人工智能融入社会和每个人的生活中。这一提议结合了一个由国际专家组成的科学委员会和一个利益相关者论坛的意见，与欧盟委员会人工智能高级专家组共同协商，为人工智能的伦理和人工智能社会的发展提出了一系列具体可行的建议。

于人们进一步关注未来几年里可能出现的特别道德问题。

下面将介绍和说明人工智能领域的原则与价值观。

- **善行原则**

人工智能系统的设计与开发应当有益于个体与集体。人工智能可促进繁荣、创造价值、增加财富、促进可持续发展。同时，致力于善行的人工智能系统有利于创造公平、包容与和平的社会，有助于维护公民的精神自治，促进经济、社会与政治机会的平等分配，从而造福全民。人工智能系统如果用于保护民主进程与法治，保障低成本与高质量的公益与服务，加强数据素养与代表性，为使用者减轻伤害及加强信任，从经济发展、社会公平与环境保护这三个方面实现联合国可持续发展目标或更广泛的可持续发展。换言之，人工智能是促进全球公益和/或应对全球挑战的有效工具。

- **无伤害原则**

人工智能系统不得伤害人类。人工智能系统的设计应确保在社会上和工作中保护人类的尊严、诚信、自由、隐私和安全。人工智能系统不得威胁民主进程、表达自由、身份自由以及拒绝人工智能服务的权利。人工智能系统至少在设计上不能加剧已经存在的伤害，也不能导致新的伤害，包括身体、心理、财务或社会等方面的伤害。人工智能领域的伤害可能来自于对个人数据的处理（包括数据的采集、存储、使用方式等）。为避免伤害，为用

于培训人工智能算法而采集和使用的数据必须要避免歧视、操纵或负面判断。同时，人工智能系统的开发与实施还必须保护社会避免意识形态的两极分化和算法决定论。

避免伤害，也包括对环境和动物的伤害。因此，避免伤害的原则应当包括开发环保型的人工智能。地球资源有其本身的价值，也有着供人类消费的价值。无论是哪一种，人工智能的研究、开发和使用有必要坚持环保意识。

- **自主原则：“保留人类的能动性”**

人工智能开发所涉及到的人类自主性，意味着人类自由不从属于人工智能，也不受其胁迫。在与人工智能系统的交互中，人类必须保持完整有效的自决权。这就要求人工智能的消费者或使用者有权决定是否遵守人工智能的直接或间接决策，有权知晓与人工智能系统进行直接或间接互动的知识，并有权选择拒绝或退出。

在许多情况下，自主决定需要政府或非政府组织确保个体或少数群体在同等条件下获得类似的机会。此外，为确保人类能动性，应当建立责任保障与问责机制。人工智能不能成为推卸人类保障基本权利的责任的借口，这一点是重中之重。

- **公正原则**

本准则所提及的公正原则要求人工智能系统的开发、使用与监管必须公平。开发者和实施者需要确保个体与少数群体享有不

遭受偏见、侮辱和歧视的自由。此外，人工智能的开发应当平衡利弊，避免加剧弱势群体的遭遇，努力为其争取在教育、商品、服务和技术上不受歧视的公平机会。公正还意味着人工智能系统必须要提供有效的救济措施，以防范意外伤害或者数据偏离人类个体或集体偏好。最后，公正原则还要求人工智能的开发者或实施者必须要拥有高标准的问责机制。为人工智能的性能制定（道德）标准，将有益于造福人类。

- **可解读原则：“透明操作”**

透明度是人工智能系统及其开发者赢得并维持信任的关键。从道德立场上来看，技术和商业模式的透明度都十分重要。技术上的透明度，是指人工智能系统要能够被理解能力和专业水平各不相同的人士所审查和理解。商业模式上的透明度，是指人工智能系统开发者与技术实施者要明确告知其意图。

可解读是个体在与人工智能交互时实现知情同意的前提，而为确保实现可解读与无伤害原则，就应当保障知情同意的要求。可解读还要求必须采取追责措施。个体与机构可要求人工智能系统的组织者与开发者、技术实施者或供应链上的第三方基本参数与说明的证据（人工智能系统的发现或预测，或者发现和预测所涉及的因素），以此为依据来分析人工智能的决策。

5、人工智能引发的重大关注

人工智能的具体用途或应用以及领域或语境可能会违背上述的权利和原则，从而引发重大关注。与其他强大的技术一样，人工智能虽然有助于实现欧盟价值观，但其两面性本质意味着人工智能也可能会侵犯这些价值观。下面列出了部分重大关注，在未来可能会有所缩减或变动。

未经同意的身份识别

公共或私营机构都可以运用人工智能来更加高效的识别个体身份。人工智能领域的控制技术必须合情合理，才能保障欧洲民众的自主权。在个体身份识别与个体跟踪之间，以及针对监控与大众监控之间进行区分，是实现可信赖的人工智能的关键。

目前互联网上也有知情同意的机制，但消费者往往不假思索就点击同意。这就要求监管者必须根据实际情况构建全新机制，让公民能够给出可验证的同意，让人工智能或同等技术能够自动识别验证。面部识别或其他使用生物计量数据的非自愿身份识别手段都属于可扩展人工智能身份识别技术（如谎言测试、通过微表情进行性格评估、自动语音识别等）。个体身份识别有时也符合道德原则（如发现欺诈、洗钱或资助恐怖分子等行为）。在现有法律或者对核心价值观的保护没有明确规定可以采用此类技术的地方，自动身份识别会引发重大的法律与道德关注，其默认条件是用户并没有给出同意身份验证的意思表示。这同时也适用

于使用能够进行逆向身份验证的“匿名”个人数据。

隐秘的人工智能系统

人工智能的开发者和部署者有责任让用户知道交互对象是人类还是机器，这一点必须可靠实现。此外，掌握人工智能的控制权，操纵人类的能力就可能达到前所未有的规模。因此，人工智能开发者和部署者应当确保人类了解交互对象是人工智能的这一事实，或者能够要求并验证这个事实。值得注意的是，存在可能让事态更加复杂的边缘情况，比如用人工智能过滤人类语音。机器人就可视为是一种隐秘的人工智能系统，因为机器人在尽量模仿人类行为。机器人进入到人类社会，有可能改变人们对人类及人性的认知。人们必须牢记，人类与机器的混合将导致多重后果，包括归属感、影响力，或者降低人类的价值。因此，类人与人形机器人的开发必须要进行谨慎的道德评估。

在未征得同意的情况下实施大规模标准化公民评分违背人们的基本权利

所有公民的自由与自主都应得到尊重。政府机构在各方面进行大规模标准化公民评分（如“道德人格”或者“道德诚实”方面的总体评估）将危及这些价值观，特别是在不符合基本权利或者规模失调且没有明确合法目标的情况下。当今世界，大小规模的公民评分常为具体领域的单纯说明性评分（如学校系统、电子

学习或驾照）。但是，公民评分如果要应用于有限的社会领域，就必须要为公民提供透明流程，并提供评分过程、目的和方法方面的信息，在理想情况下还要向公民提供退出评分机制的选择权。在双方权利不对称的情况下，这一点尤为重要。因此，开发者和部署者都应当在技术设计中保障这种选择权，并提供必要的资源予以实现。

致命自动化武器系统（LAWS）

致命自动化武器系统可以在没有人类控制的情况下自动执行选择与攻击个体目标的关键功能。而最终，所有伤亡还是必须要由人类负责。目前，不明数量的企业和国家正在研究和开发致命自动化武器系统。这种情况会引发基本的道德关注，比如会导致前所未有的军备竞赛失控，会导致人类完全放弃控制、故障风险完全无解的军事环境。不过还有另一方面需要注意，那就是致命自动化武器系统能够在武装冲突中减少附带伤害，比如有选择的解救儿童。欧洲议会呼吁紧急制定共同遵守的法律，解决人类控制、监督的道德与法律问题，落实国际人权法律、国际人道主义法律和军事战略的问责与落实机制。

远期关注

当前所有的人工智能仍然仅涉及具体领域，需要训练有素的人类科学家和工程师为其指定准确目标。但是，长远的未来仍然

存在一些重大关注，虽然这些仅仅是猜测。尽管以现在的眼光看来，出现这些情况的可能性并不大，但其潜在危害却可能十分巨大（比如拥有主观经验的人工意识、人工道德实体或者无监督递归自我改进型人工智能（AGI）等，这些技术在今天看来尚遥遥无期）。因此，人们需要采取风险评估策略，始终关注这些领域的发展，并投入资源减少对远期风险、未知事件以及“黑天鹅”事件的认知空白。并欢迎参与讨论的有识之士踊跃建言。

确保道德目标的主要指导原则

- 确保人工智能应以人为本：人工智能的开发、部署和使用必须要符合上述的“道德目标”，要遵循基本权利、社会价值观以及善行、无伤害、人类自主性、公正和可解读的道德原则。这是实现可信赖的人工智能的关键所在。
- 基于基本权利、道德原则和价值观，以前瞻性的眼光评估人工智能可能给人类及公益带来的影响。要特别关注涉及儿童、残疾人、少数群体等弱势群体的情况，以及劳资或企业与消费者等双方权力或信息不对等的情况。
- 要认清人工智能利弊并存的事实，必须对存在重大关注的领域保持警惕。

（二）实现可信赖的人工智能的要求

本部分提供了落实和实现可信赖的人工智能的指导原则。阐

述了可信赖的人工智能的主要要求，确保人工智能的发展能够扬长避短。

实现可信赖的人工智能，意味着要将抽象的总体原则具体落实到人工智能的系统和应用之中。本文根据第一部分提出的权利、原则和价值观推导出了以下 10 点要求。

- 问责
- 数据管理
- 大众化设计
- 对人工智能自主性管理（人类监督）
- 非歧视
- 尊重（并加强）人类自主性
- 尊重隐私
- 可靠
- 安全
- 透明度

上述要求并未穷尽列举，所给出的可信赖的人工智能的要求按字母表顺序排列，必须强调所有要求均同等重要。在第三部分里，提供了落实这些要求的评估内容。

（三）可信赖的人工智能的评估

本部分旨在对上述可信赖的人工智能的各项要求进行可操作

的具体落实与评估，涵盖人工智能开发与使用各个阶段。本文建议使用评估内容来指导开发者、部署者和其他创新人员实现道德目标和技术可靠。

评估内容提出了评估人员需要思考的问题。这些问题并非穷尽，而只是初期制定。而且，各种情况面临的具体问题各不相同，需要针对人工智能所面临的具体情况专案专办。

应当注意，基于评估内容的评估并非孤立措施，必须要与涵盖人工智能道德框架的管理流程相结合。

有助于实现可信赖的人工智能要求的评估内容如下:

1、问责

- 出错时由谁负责?
- 是否具备承担责任的技能和知识?（负责任的人工智能培训?道德宣誓?）
- 第三方或员工能否上报潜在的漏洞、风险或偏见，是否有处理这些问题和报告的流程?这些流程是否都有单一的联系点?
- 是否预见到人工智能系统的（外部）审核?
- 人工智能系统的相关人员招聘是否为确保背景多样化而实施了多样性和包容性的政策?
- 是否成立了人工智能道德评估委员会?是否建立了讨论灰

色领域的机制？是否有内部或外部专家组？

2、数据管理

- 是否有保障适当的数据与流程管理？为保障适当的数据管理而采取了什么流程？
- 是否有监督机制？由谁最终负责？
- 人工智能系统适用哪些数据管理条例与法规？

3、大众化设计

- 系统应用是否平等？
- 系统是否适应广泛的个体偏好和能力？
- 有特别需求或有残疾的人士是否可使用系统，这方面采用了何种设计和验证途径？
- 系统开发和/或部署适用的公平定义是什么？
- 适用的平等措施要如何衡量和保障？

4、对人工智能的自主性管理

- 每个阶段有无在必要时进行人类控制的流程？
- 自我学习人工智能中有无“停止按钮”？规范性（自主决策型）人工智能有无此按钮？
- 人工智能系统在什么情况下可视为具备无需人类监督或控制的自主能力？
- 为确保人类承担人工智能系统决策的全部责任而采取了哪

些措施？

- 对人工智能自主性管理的相关问题采取了哪些审核与补救措施？
- 机构内部由谁负责核实人工智能系统是否能够适当管理、并由人类负责其行为？

5、非歧视

- 有哪些因素会导致相同执行条件下产生不同决策？这种差异会不会影响基本权利或道德原则？应如何衡量？
- 在差异之间进行取舍时有无明显的偏见？
- 有无避免形成或避免加剧数据与算法偏见的策略？
- 在系统开发与使用阶段有无针对这种偏见进行持续测试的流程？
- 有无向对象或群体清楚解释与歧视相关的问题，特别是人工智能系统的使用者或受到其影响的其他实体遇到的歧视问题？

6、尊重隐私

- 系统是否符合《通用数据保护条例》的相关规定？
- 系统中的个人数据信息是否符合现行隐私保护法律？
- 使用者要如何找到有效表示同意的信息，以及如何撤回同意的信息？
- 有无向对象或群体清楚解释侵犯隐私的相关问题，特别是

人工智能系统的使用者或受到其影响的实体遇到的侵犯隐私问题？

7、尊重（并加强）人类的自主性

- 有无告知使用者产品存在影响人类精神完整性的风险？
- 是否向用户提供了服务/产品的必要信息，使其能够做出完全自主的决策？
- 人工智能系统有无告知使用者决策、内容、建议或结果是算法决策的结果？
- 使用者有无为知晓算法目的、来源及依据数据等信息而质问算法决策的措施？

8、可靠

应对攻击的弹性能力：

- 人工智能系统容易遭受什么样的攻击？哪些攻击可以抵消？
- 哪些系统能够确保数据安全与完整性？

可靠性与再现性：

- 有无监控测试产品或服务是否符合目标及预期应用的策略？
- 所使用的算法是否进行了再现性测试？再现性的条件是否可控？在哪些具体和敏感条件下有必要使用不同的方法？
- 对于应当考虑的可靠性与再现性方面，要如何予以衡量和保障？

- 负责人工智能系统开发与测试的实体有无人工智能系统可靠性测试与验证流程文件和实用指导原则？
- 有哪些机制可为使用者确保人工智能系统的可靠性？

数据使用与控制的准确性：

- 所研究和/或使用的系统适用准确性定义是什么？
- 为考虑各种形式的准确性，应如何予以衡量和保障？
- 数据是否足够全面，能够完成现有的任务？是否使用了最新的数据（非过时的数据）？
- 还有哪些数据源/模型能够提高准确性？
- 还有哪些数据源/模型可用于消除偏见？
- 采用了什么策略来衡量数据的包容性？数据是否足以代表目标案例？

后备计划：

- 人工智能系统的以下错误会产生什么影响：提供错误结果；无法使用；提供社会无法接受的结果（如偏见）？
- 是否针对上述情况（产生不可接受的影响时）确定了后备计划的触发条件？
- 后备计划是否已经确定并测试？

9、安全

- 所研究和/或部署的系统相关的安全定义是什么？

- 各种形式的安全要如何衡量和保障?
- 是否已找出可预见技术应用的潜在安全风险，包括意外或恶意滥用的风险?
- 在出现人身完整性风险时是否提供信息?
- 使用产品或服务的相关潜在风险是否有分类与评估的流程?
- 是否已制定抵消和/或管理已确定风险的计划?

10、透明度

目标:

- 产品/服务是否有清楚的获益对象?
- 是否已确定并清晰传达产品的使用情境?
- 是否已向使用者说明产品的局限?
- 产品是否已制定部署标准并提供给使用者?

可追溯性:

- 已经采取了哪些确保产品准确性的措施?
- 产品或技术的性质及其潜在或已知风险（如偏见方面）是否已经能够让预期使用者、第三方和公众所访问和理解?
- 人工智能系统是否拥有在关键情况下保障可审核性的追溯机制？这要求应有以下规定:
  - 构建算法系统的方法
    - 如果是规则型人工智能系统，应当能够澄清人工智能

系统的编程方法（如模型构建方式）；

- 如果是学习型人工智能系统，应当能够澄清培训算法的方法。这就需要提供所用数据的信息，包括：所用数据的采集方式；所用数据的选择方式（如采用了何种择选或排除标准）；培训算法时是否使用了个人数据？请注明使用了哪种类型的个人数据。

➢ 算法系统的测试方法

- 如果是规则型人工智能系统，应当提供测试及验证系统的情境选择或测试案例；
- 如果是学习型模型，应当提供系统测试所用数据的相关信息，包括：所用数据的采集方式；所用数据的选择方式；培训算法时是否使用了个人数据？请注明使用了哪种类型的个人数据。

➢ 算法系统的结果

- 应当提供算法决策的结果，以及不同情况下可能产生的其他决策（如其他分组）。

备注：如上所述，由于人工智能有着具体的应用，任何评估都必须针对人工智能系统的具体应用加以定制。为实现评估内容的具体操作性，本文参考了 52 位人工智能高级专家组专家以及欧洲人工智能联盟成员的意见，进而总结了四个具体应用案例：

（1）医疗诊断与治疗；（2）自动驾驶/出行；（3）保险理赔；（4）资料收集和执法。

可信赖的人工智能的评估的主要指导原则

- 在人工智能的开发、部署或使用阶段采用可信赖的人工智能的评估内容，并根据具体情况对评估内容进行调整。
- 应牢记评估内容不可能穷尽所有情况，而且实现可信赖的人工智能不能只是照单打勾，而是在人工智能系统的整个生命周期不断确认要求、评估方案、确保提高改进的持续过程。

三、结论

本文为人工智能高级专家组制定的《人工智能道德准则草案》。人工智能高级专家组认为，人工智能已给全球商业和社会带来巨大的积极影响。人工智能是一种变革与颠覆性技术。过去几年，得益于海量数字化数据、计算能力与存储能力的重大进步以及人工智能方法与工具的重大科研创新，人工智能实现了跨越式发展。人工智能将继续给社会和民众带来难以预料的影响。这种前景令人振奋，但人们也必须保持警惕，应正确认识可信赖的人工智能，并努力加以践行，只有赢得人类的信任，技术才能发挥造福人类

的巨大作用。因此，实现可信赖的人工智能是本文件的根本宗旨。

可信赖的人工智能包含两个方面：第一，应尊重基本权利、相关规则和核心原则，确保“道德目标”的实现；第二，技术上必须稳健可靠。然而，使用人工智能也可能带来意外伤害。因此，人工智能高级专家组独辟蹊径，制定了实现可信赖的人工智能的框架和具体指导原则。

人工智能高级专家组欢迎各相关方通过欧洲人工智能联盟，就本草案提出相关意见和建议。人们应清醒认识到，本草案仅代表人工智能专家组目前的研究成果，因此仅适用于当前情况。

始终坚持以人为本的传统使欧洲能够高瞻远瞩。以人为本已经深深植根于欧盟的各项条约，成为欧洲的基因。本文同样秉承以人为本的人工智能发展策略，旨在使欧洲在坚持道德目标的前提下，成为引领全球人工智能发展的创新者。这一宏大愿景有助于在全体欧洲民众中掀起一股热潮，本准则的目标是打造一种文化，即“可信赖的人工智能源自欧洲”。

译自：*Draft Ethics Guidelines for Trustworthy AI, 18 December 2018 by the European Commission*

译文作者：工业和信息化部赛迪研究院　席子祺
联系方式：18500989079
电子邮件：xiziqi@ccidthinktank.com

赛迪智库
面向政府 服务决策

研究，还是研究
才使我们见微知著

规划研究所	知识产权研究所	安全产业研究所
工业经济研究所	世界工业研究所	网络安全研究所
电子信息研究所	无线电管理研究所	中小企业研究所
集成电路研究所	信息化与软件产业研究所	节能与环保研究所
产业政策研究所	军民融合研究所	材料工业研究所
科技与标准研究所	政策法规研究所	消费品工业研究所

编 辑 部：工业和信息化部赛迪研究院
通讯地址：北京市海淀区万寿路27号院8号楼12层
邮政编码：100846
联 系 人：王 乐
联系电话：010-68200552 13701083941
传 真：010-68209616
网 址：www.ccidwise.com
电子邮件：wangle@ccidgroup.com

报：部领导

送：部机关各司局，各地方工业和信息化主管部门，
相关部门及研究单位，相关行业协会

编 辑 部：工业和信息化部赛迪研究院
通讯地址：北京市海淀区紫竹院路 66 号赛迪大厦 15 层国际合作处
邮政编码：100048
联 系 人：姚 丹
联系电话：（010）88559684　13811086893
传　　真：（010）88558833
网　　址：www.ccidgroup.com
电子邮件：yaodan@ccidgroup.com