2021世界人工智能大会的主题论坛之一，由世界人工智能大会组委会办公室主办，中国信息通信研究院（以下简称“中国信通院”）与京东探索研究院联合承办的2021世界人工智能大会-可信AI论坛（以下简称“可信AI论坛”）于2021年7月9日在上海浦东世博中心金厅召开。

论坛现场，澳大利亚科学院院士、京东探索研究院院长陶大程博士发表了题为《可信人工智能的前世今生》的主题演讲，结合刚刚发布的国内首本《可信人工智能白皮书》，全面地回顾了可信人工智能的提出背景和发展历程，就实现可信AI的实现路径及稳定性、可解释性、隐私保护、公平性等研究方面进行了系统阐述，就新一代人工智能的未来进行展望。



Marked：陶大程博士现场发表演讲

随着时代的快速发展，人工智能技术在许多垂直领域已有广泛应用，如图像识别、内容推送、自动翻译等。然而，人工智能技术在应用的同时，也面临许多可信挑战。例如，特殊构造的图案会误导商品识别系统产生识别错误；缺乏可解释性的自动驾驶系统可能引发人们对人工智能技术的信任危机。

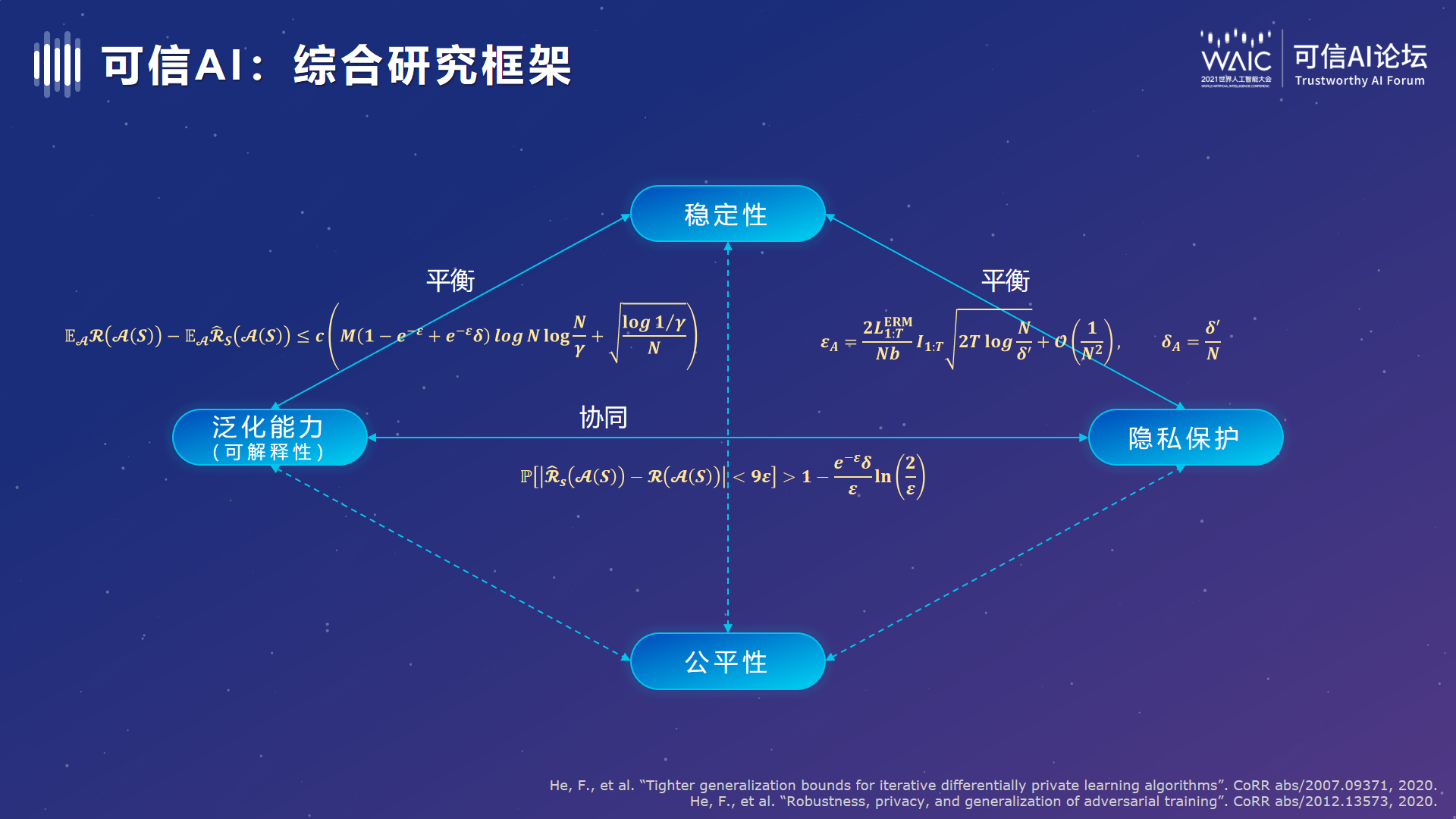
据陶大程博士介绍，人工智能技术的可信挑战引起了国内外的广泛关注。国际上，欧盟、美国、澳大利亚等世界主要国家及组织纷纷提出各自的可信人工智能治理计划。在国内，国家新一代人工智能治理专业委员会于2019年6月发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了人工智能治理的框架和行动指南；腾讯、京东等科技公司也在如何应对可信挑战方面展开了诸多探索。

例如，2019年10月，京东集团就首次在乌镇世界互联网大会上提出京东践行“可信赖AI”的六大维度；京东探索研究院在2021年4月已将“可信人工智能”正式列为主要研究方向之一，并联合中国信通院完成撰写、发布国内首本《可信人工智能白皮书》。

陶大程博士认为，可信人工智能技术是帮助人工智能实现可信的基础。从技术角度来说，可信人工智能的研究可以归结为稳定性、可解释性、隐私保护及公平性四个方面，从以上四点出发研究可信人工智能，首要任务是找到合适的方法来定量分析、量化人工智能算法、模型和系统在稳定性、可解释性、隐私保护及公平性方面的能力。

目前，稳定性可以通过各类攻击算法的攻击成功率或攻击性能来衡量；可解释性中的泛化能力，可以通过泛化误差上界来描述；隐私保护能力常通过差分隐私法来刻画，也能通过各类隐私攻击算法的攻击成功率或攻击性能来评估；公平性则使用一系列公平决策指标来进行衡量，这些指标主要分为个体公平性及群体公平性两大类。

同时他强调，这四个方面内部存在着深层次联系，彼此之间并不孤立。因此，在开展可信人工智能研究时，不能仅从单一维度出发，而需要从整体综合考虑不同因素之间的影响。这样，才能更好地进行相关研究，从而推动可信人工智能的进一步发展。



展望可信AI的未来，陶大程博士预测未来可信AI一体化研究将是重要趋势，而可信能力评测也将成为人工智能产业落地过程中的重要一环。从理论与实践层面持续开展可信AI研究，将推动人工智能产业新浪潮，帮助人工智能实现可信，将充分释放人工智能的潜力，并更好地造福人类社会。

目前，可信人工智能已作为京东探索研究院锁定的三大研究方向之一，京东探索研究院致力于从基础理论层面实现颠覆式创新，助力数智化产业发展及变革，集合京东集团资源和能力，专注前沿科技探索，致力于实现研究和协同创新的生态平台，并以原创性科技赋能零售、物流、健康、科技等全产业链场景，打造源头性科技高地。