我们是否能够有意义的评估那些操作不透明的系统的活动？

首先，我们是可以在一定程度上有意义的评估那些操作不透明的系统的活动，尽管这是挑战。

在我们评估那些操作不透明的系统活动问题时，我们需要采用一些特殊的方法，在‘Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour‘一文中，作者提出了两种方法：摩根准则和休谟准则。 其中摩根准则是优先解释非人类行为为简单的心理过程反应，而不是更复杂的治理过程。这可以直观体现摩根准则的关键点是谨慎和简洁，我们最不愿看到的就是将人工智能行为进行过度解读，而摩根准则就有很好的帮助来避免我们引入了不必要的复杂性，是我们更能理解AI系统的实际功能和局限性。

可解释性人工智能（XAI）在这里的作用是什么？它是否解决或绕过了“不透明性”问题？

然而在深度学习的研究中存在比较偏见，包括人类中心主义，人工智能人格化，选择性比较等。人类是需要在黑盒子AI模型中识别和解决偏见，所以透明度在AI模型中是非常重要的。但是黑盒子AI模型的算法透明度不足，这使得在这些模型中识别和解决偏见变得更加困难。因此，可解释性人工智能（XAI）就可以很好的帮助提高AI模型的透明度和可解释性。

在一定程度上，XAI具有局限性，它只能检测其简化假设所代表的偏见，而且只有当技术专家能够通过反思意识来改变其简化假设的选择时，XAI模型才能够发挥作用。此外，相关的解释性资源通常没有纳入主流认知资源，因此在实践中，技术专家很难获得这些资源。此外，由于技术专家相对独立于各种利益相关者的工作，这种缺乏互动进一步阻止了他们适当地调整其简化假设的选择。 所以，在‘Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy‘ 一文中，为了开发更综合的可解释性人工智能（XAI），作者主张需要吸引各种利益相关者参与同行评审过程，并融入女性主义哲学，以确保XAI的开发更加透明、负责和公平。女性主义哲学在解决算法偏见问题中发挥着关键作用，特别是关注社会正义和公平性。它强调知识的社会性和局限性，主张知识的生产和传播应更加包容和多元化，以避免忽视或排除边缘化群体的知识。通过融入女性主义哲学，我们能更好地理解算法偏见问题的社会和文化背景，确保算法的开发和应用更加公正和负责。此外，女性主义哲学提供了一些有用的工具和方法，如培养认识论美德和建立认识论民主，以助解决算法偏见问题中的知识不对称和权力不平等问题。

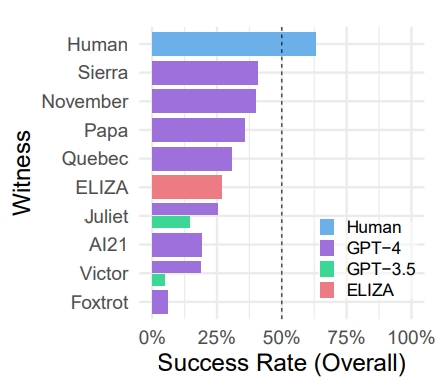
客观的说，可解释人工智能（XAI）并没有完全解决或者绕过人工智能系统中的不透明性问题。尽管XAI可以一定程度上帮助我们更好的理解人工智能系统的决策过程，但是它只包含最少量的客观信息。这意味着XAI提供的信息有限，也仅仅能帮助我们了解人工智能系统是如何做出决策的，但并没有涵盖到可能影响决策的所有价值观和假设。正如‘Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy‘一文中作者提供的例子，作者以一个不透明的人工智能招聘系统为例，该系统采用深度神经网络模型来预测求职者的’录取可能性分数‘，该分数基于一百多个特征，比如年龄、性别、种族、个性等特征。但是，作者认为XAI的模型并非通用，更广泛的意义构建和知识产生资源是必要的，其中包括关于常常是偏见和压迫受害者的边缘群体的深刻社会知识。所以XAI模型是对复杂人工智能系统的有意见话表示，因此无法准确代表左右方面。因此，XAI模型只能准确检测有限范围的偏见，其表征目标的一些特征被抽象，同时一些不切实际的特征被合理化。总之，作者认为仅依赖技术XAI无法解决人工智能系统不透明性问题，需要采取更全面的方法，如综合XAI，以确保人工智能系统是透明、可问责和公平的。

AI系统是否以人类易懂的民间心理词汇解释是多么重要？特定任务是否有影响？

总的来说，使用人类能理解的语言是一个非常复杂的问题，答案可能取决于具体的任务。但是这是十分重要的。在文章‘Dual-Process Theories of Higher Cognition: Advancing the Debate‘中提到在需要像非专业人士结束AI系统的决策树过程时，使用人类易懂的民间心理词汇解释是非常重要的，这种解释可以提高人们对AI系统的信任度和可靠性，并使人们更容易理解AI系统的工作原理。这也可以很好的拉近AI与非专业人士的距离，使他们更好地了解和使用AI系统。

在特定任务时例如金融，司法，医疗等情况下使用专业化的术语和解释是非常有必要的，因为在这些领域，确保人工智能系统是透明且可解释是至关重要的，这可以保证信息传达的准确性和效率。而且依旧可以使用户可以信任系统并对其决策负责。然而在客户服务等服务型任务情况下可能需要更加友好和耐心的语言格。而且，必要的解释水平可能因具体任务和背景而异。在某些情境下，对模型输出进行简要解释可能足够，而在其他情境下，可能需要更为详细的解释。总体而言，可解释人工智能的目标是向用户提供必要的信息，使其能够明智地决策如何使用和信任人工智能系统。

当我们提到人类易懂的民间心理词汇时，图灵测试是需要被探讨的。在”Does GPT-4 Pass the Turing Test?” 一文中指出人工智能模型的语言风格和社交情感特征是评估其人类相似度的重要因素之一。这表明，使用人类易懂的心理词汇解释能力是评估人工智能模型在自然交流中性能的因素之一。在文中作者进行了一场公开的在线图灵测试，参与者在聊天界面中与人工智能代理和人类参与者进行互动。参与者被随机分配为审问者或证人的角色。 审问者的目标是弄清楚证人是人类还是人工智能，而证人的目标是说服审问者他们是人类。结果如下图所示。



*Figure 1: Overall Turing Test Success Rate (SR) for*

*a subset of witnesses* [4]

在这项实验中我们可以看到，人类的成功率是63%，目前是表现最好的。而GPT-4以最好的成功率41%（Sierra）通过了测试。虽然没有超过人类的成功率，但这看似也是一个不错的结果。在文章‘Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT‘ 中指出ChatGPT模型依赖于思维链推理，使它们能够制定解决任务所需的策略，检查起始假设，估计部分解决方案或测试替代方法。因此，ChatGPT模型正确回答并避免直觉偏见的能力源于它们独特的响应风格，这依赖于思维链推理。这是一种推理方式，涉及使用输入输出上下文窗口制定解决任务所需的策略、检查起始假设、估计部分解决方案或测试替代方法。这类似于人们如何使用记事本解决数学问题或写文章来发展他们的论点。

在这篇文章中还指出，将心理学方法应用于研究大型语言模型。因为这可以揭示先前未被察觉的新兴特性。最近的研究发现，随着LLMs复杂性的增加，它们展现出多种技能和特性，其中一些并非其创建者先前预期或意图。这一发现对于这些模型在各种应用中的开发和部署具有重要意义，如自然语言处理、聊天机器人和虚拟助手。通过了解这些模型的局限性和偏见，我们可以更好地设计和评估它们的性能，并确保它们以负责任和道德的方式使用，这也使人工智能系统在特定任务中获得更好的表现。在文章‘The Ethics of Terminology: Can We Use Human Terms to Describe AI?‘，中有一个例子很好的很好的说明了这个情况。在文章中，作者提到了一个“文化匹配”的实用论证，即使用熟悉的类别来传达复杂的技术概念是可以被接受的，即使这样做会失去一定的精度。例如，使用“蘑菇”一词来描述真菌丝，因为这个类别对人们来说是熟悉的，而人们可能不熟悉真菌丝的区别或技术术语。这种方法可以使官方公共话语对公民来说更易于理解或接受，从而提高传播效率和接受度。 文章还提到了其他类似的例子，例如在法律文本中使用与公民直觉相符的术语可以使法律文本更有效。这种方法可以使复杂的技术概念更易于理解和接受，从而提高传播效率和接受度。然而，这种方法可能会失去一定的精度，因为使用熟悉的类别来传达复杂的技术概念可能会导致一些精度上的损失。因此，这种方法只有在传播效率和接受度得到提高，并且最终的传播结果对接收者有益时才是合理的。

批判性的说，在‘Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy‘中，作者专注于女性主义哲学可能并非解决算法偏见的唯一途径。尽管作者为在知识生产中纳入多元化的观点和经验提出了令人信服的论点，但也有人认为其他方法，比如增加开发团队的多样性或使用更多元化的训练数据，同样可能有效地减轻算法偏见。此外作者关于在人工智能开发中需要纳入女性主义哲学的论点在所有情境中可能并非可行或切实可行。有人可能认为纳入女性主义哲学可能需要在开发过程中进行重大变革，这可能耗时且昂贵。此外，一些人可能认为在某些领域或情境中，如军事或安全应用，纳入女性主义哲学可能不相关或不适用，其他因素，如国家安全，可能优先于公平和透明度。因此，尽管作者的论点很有说服力，但在所有情境中可能并非普遍适用或可行。还有图灵测试可能是一种过时而且对当代来说不完善的测试方式，因为图灵测试过于依赖语言风格和社会情感特征，而不是实际的智能和推理能力。而且这没有一个明确且统一的标准参考，所以结果可能不是十分可靠。此外，该研究未考虑到Ai模型能够冒充人类并欺骗人们可能带来的潜在道德影响。

最终，根据上文所讨论的内容在评估操作不透明的系统活动时，采用特殊方法能够有意义的进行评估，例如摩根准则。虽然可解释性人工智能（XAI）在提高透明度方面发挥作用，但其局限性需要通过吸引各种利益相关者参与同行评审、融入女性主义哲学来解决。然而，XAI并不能完全解决系统不透明性问题，需要更全面的方法，如综合XAI。在使用人类易懂的民间心理词汇解释方面，平衡专业术语和友好语言取决于任务性质。在特定任务中，如金融和司法领域，使用专业术语，这是十分重要的。女性主义哲学可能并非解决算法偏见的唯一途径，其他方法也可能有效。图灵测试的局限性表明需要更全面、考虑道德影响的评估方法。