人工智能帮你测谎？！当心强化偏见

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/27920.html

人工智能帮你测谎？！当心强化偏见。 虽然人类经常说谎，但通常不会指责别人说谎，因为社会规范禁止虚假指控，并强调文明礼貌。但人工智能（AI）可能很快就会改变这些规则。德国科学家研究发现，当人工智能提出指控时，人们更有可能指责他人撒谎。相关研究6月27日发表于《交叉科学》。

我们的社会对撒谎指控有着牢固而完善的规范。论文通讯作者、杜伊斯堡-
埃森大学行为科学家Nils Kobis说，公开指责别人撒谎需要很大的勇气和坚实的证据。但我们的研究表明，人工智能可能成为人们避免为指控的后果负责的借口。

人工智能测谎带来更多指责

人类社会长期以来都是基于默认真话理论来运作的，该理论认为，人们通常认为他们听到的都是真的。为此，人类很难发现谎言。此前的研究表明，人们在识破谎言方面的表现并不比随机监测更好。

Kobis和团队想知道人工智能的存在是否会改变这些既定社会规范和行为。

在调查中，研究小组要求986名参与者对他们下周末的计划写一份真实和一份虚假的描述。该团队用这些数据训练了一种算法，开发出一种人工智能模型，该模型识别真假陈述的准确率能达到66%，这明显高于普通人的水平。

接下来，该团队招募了2000多人作为评委，他们将阅读一份陈述，并决定它的真假。研究人员将参与者分为四组——基线、强迫、封闭、选择。

在基线组中，参与者在没有人工智能帮助的情况下，回答对或错。在强迫组中，参与者在做出自己的判断之前总是收到AI预测。在封锁组和选择组中，参与者可以选择是否接受人工智能生成的预测。但封锁组的人不会收到预测反馈，而选择组的人会收到。

研究小组发现，基线组的参与者在判断真假时的准确率为46%。只有19%的人表示他们读到的陈述是假的，尽管他们知道50%的陈述是假的。这证实了人们倾向于避免指责别人撒谎。

在强迫组中，不管参与者是否想要，他们都得到了人工智能预测结果，因此超过1/3的参与者认为陈述是虚假的。这一比例明显高于基线组和未接受人工智能预测的组。

当人工智能预测某句话是真的时，只有13%的参与者认为这句话是假的。然而，当人工智能预测一个陈述为假时，超过40%的参与者认为该陈述是假的。

此外，在要求并收到人工智能预测的参与者中，84%的人接受了预测结果，并在人工智能说陈述是错误的时候提出了指控。

这表明，一旦人们手头有了这样的算法，他们就会依赖它，并可能改变他们的行为。如果算法认为某件事是谎言，人们就会欣然接受。这是相当令人担忧的，这表明我们应该非常谨慎地对待相关技术。Kobis说。

当心人工智能强化偏见

但有趣的是，人们似乎不愿意使用人工智能作为测谎工具。

在封锁组和选择组中，只有1/3的参与者选择接受人工智能预测。这个结果让研究小组感到惊讶，因为研究人员事先告诉参与者，该算法比人类更能发现谎言。

这可能是因为人们对自己的测谎能力过于自信，尽管人类在这方面真的很糟糕。Kobis说。

研究人员认为，这一发现有助于深入了解使用人工智能系统进行测谎的社会影响，并在实施类似技术时为政策制定者提供信息。

人工智能以频繁犯错和强化偏见而闻名。鉴于这些发现，Kobis建议政策制定者应该重新考虑在重要而敏感的问题上使用这项技术，比如庇护批准等。

人工智能被大肆宣传，许多人认为这些算法非常强大，甚至是客观的。我很担心这会让人们过度依赖它，即使它的效果并不好。Kobis说。（来源：中国科学报 冯维维）

相关论文信息：http://doi.org/10.1016/j.isci.2024.110201

作者：Nils Kobis 来源：《交叉科学》