该文件探讨了利用强化学习（RL）开发自动做市策略的问题。所面临的挑战包括处理有毒订单流和库存风险，这可能导致在与信息更灵通的交易对手进行交易时出现逆向选择，或在趋势市场中积累不希望出现的头寸。

所使用的方法包括模拟市场对做市商行为的反应。这需要使用RL来利用环境（模拟器）中的弱点。该文件还提到在学习做市时使用近端策略优化（PPO）。作者将学习到的策略与最优策略进行了比较。他们还提到了Álvaro Cartea、Sebastian Jaimungal和José Penalva所著《算法和高频交易》一书中数学模型的使用。

结果包括开发了一种能够有效模拟市场反应的做市商策略。作者还介绍了一种学习型PPO代理，并将其与CarteaJaimungal代理进行了比较。

该研究的局限性包括需要在交互真实性方面开展更多工作。虽然他们分析的条件生成对抗网络（CGAN）具有令人印象深刻的孤立真实性，但他们能够构建手工制作的基于RL的交易代理，利用其特征和机制赚取不切实际的利润。这表明在交互真实性方面还需要做更多的工作。该文件还提到，在训练具有大量经验的良好策略时，RL的样本效率较低。此外，该文件还强调，通过实时交易从零开始建立RL通常是不可行的，因此需要使用某种形式的模拟环境。

通过实盘交易从头开始强化学习（RL）通常被认为是不可行的，原因有以下几点：

高风险：RL算法从试验和错误中学习。在实盘交易环境中，这意味着算法很可能会犯很多错误，并在学习到有利可图的策略之前遭受重大损失。

数据效率： RL算法通常数据效率较低，这意味着它们需要大量数据（经验）才能有效学习。在实时交易环境中，获取如此多的数据可能需要很长时间。

非平稳性： 金融市场是非稳态的，这意味着其动态会随时间而变化。今天训练的RL算法在未来可能表现不佳，因为它所训练的市场条件可能不再适用。

成本： 实时环境中的每笔交易都需要支付交易费用。从零开始学习的RL算法可能会执行大量的交易，这可能会导致非常高的成本。

监管和合规问题： 金融市场受到严格监管。通过实时交易学习的RL算法可能会违反交易规则或法规，从而导致罚款或其他处罚。

出于这些原因，在模拟环境中训练RL算法通常更为实际，在模拟环境中，RL算法可以从历史数据中学习，犯错误而不会产生实际后果，并在各种市场条件下进行测试。