WrittenAssignment_111652026劉馥瑞

1.

SDE 的機率密度 p(x,t)滿足 Fokker–Planck 方程：

$\partial p(x,t)/\partial t = -\partial/\partial x\,[f(x,t)p(x,t)] + (1/2)\,\partial^2/\partial x^2\,[g^2(x,t)p(x,t)]$ -(1)

我們希望找到一個ODE：

$dx_t/dt = v(x_t, t)$

使得此 ODE 所生成的密度流動與上面的隨機過程 x_t 產生的密度演化相同。

若密度 p(x,t) 滿足連續性方程：

$\partial p/\partial t = -\partial/\partial x\,[v(x,t)p(x,t)]$ -(2)

令(1)(2)右式相等：

$-\partial/\partial x\,[vp] = -\partial/\partial x\,[fp] + (1/2)\,\partial^2/\partial x^2\,[g^2p]$

等號左右積分一次：

$v(x,t)p(x,t) = f(x,t)p(x,t) - (1/2)\,\partial/\partial x\,[g^2(x,t)p(x,t)]$

化簡：

$v(x,t) = f(x,t) - (1/2p)\,\partial/\partial x\,[g^2p] = f(x,t) - (1/2)\,\partial g^2/\partial x - (1/2)\,g^2\,\partial \log p/\partial x$

因此，Probability Flow ODE 為：

$dx_t = [\, f(x_t,t) - (1/2)\,\partial/\partial x\, g^2(x_t,t) - (1/2)\, g^2(x_t,t)\,\partial/\partial x \log p(x_t,t)\,]\, dt$

2.

(a)

我認為二十年後，人工智慧最重要的突破將是——能自主進行倫理推理與道德決策的 AI 系統。目前的 AI 只能根據人類預設規則或偏好做出選擇，缺乏「價值推理」的能力。然而，隨著 AI 深度參與醫療、司法、教育與政治等領域，未來的社會將迫切需要一種能「理解人類道德複雜性」的智能。

想像一個情境：自動駕駛車即將發生意外，AI 必須在幾毫秒內判斷行動方案——是保護乘客，還是避開路人？或在醫療分配中，AI 需決定哪位病患先接受急救資源。現今的系統只能依照程式規則行事，但未來的 AI 能透過道德推理框架，分析情境、評估後果、對比社會價值，最終提出可被人類理解並接受的決策理由。

這樣的能力將讓 AI 不僅是工具，而是社會參與者。AI 將能在倫理審查、法律判例、公共政策等領域提供具理性與道德一致性的建議，協助人類處理極端或模糊的價值衝突。這不代表讓機器取代人類道德，而是讓 AI 成為「價值思考的助理」，協助我們反思決策後果、揭示偏見與矛盾。

(b)

要讓 AI 具備倫理推理能力，將結合以下幾類學習：

- 監督式學習：透過大量人類倫理判斷資料（如哲學實驗、法律案例）學習人類的道德模式。
- 非監督式學習：從跨文化文本（文學、歷史、宗教經典）中學習潛在的價值語意結構。
- 強化學習：AI 在模擬社會環境中嘗試行動，根據人類回饋（例如「此行為被認為正確或錯誤」）不斷更新行為策略。

在這個框架中，「資料來源」是人類行為與語言中蘊含的倫理決策紀錄；「目標訊號」則是「社會可接受性」與「道德一致性」的程度。AI 的學習過程中包含人類互動回饋，形成持續調整的道德認知體系。

(c)

第一階段的「模型化」

為了邁向具倫理推理能力的 AI，第一個研究步驟可設計為：

「讓 AI 學會在模擬的道德困境情境中解釋自己的選擇理由」

例如，建立一個「道德決策模擬平台」，提供各式倫理兩難情境（如「電車難題」、「醫療分配」、「環境取捨」），AI 必須選擇行動並生成可理解的理由。

- 概念代表性：這個簡化問題正是未來倫理推理 AI 的縮影——在價值衝突下仍能理性解釋決策。
- 可測試性：可透過人類調查評估 AI 回答的倫理合理性與一致性。
- 所需工具：大型語言模型（LLM）、因果推理、可解釋強化學習（Explainable RL）與價值對齊理論（Value Alignment）。

3.

When we forward both images to time T (so they become approximately distributed as $N(0,\sigma_T\text{^}2I)$and linearly interpolate, does this interpolation truly preserve *semantic* continuity? What does linear interpolation mean geometrically in a high-dimensional Gaussian space?