จากผล Review ของ PPO_V2 ผมมองว่าปัญหาหลัก **ไม่ได้อยู่ที่ PPO algorithm** แต่เป็นเรื่อง **Objective Design + Market Representation + Learning Problem Definition**

ตอนนี้ PPO_V2 เรียนรู้สิ่งที่ถูกต้องตาม Reward ที่ให้มัน:

> "รักษาเงิน ลด drawdown และถือ cash เมื่อไม่มั่นใจ"

ซึ่งมันทำได้ดี

แต่เป้าหมายใหม่ของคุณคือ:

> "Agent ต้องเรียนรู้เพื่อสร้าง Alpha เอาชนะตลาดในทุก regime, Sharpe > 1.5-2, Maximum Drawdown ต่ำ"

นั่นคือโจทย์คนละระดับ

เราต้องเปลี่ยนจาก:

```
Risk Avoidance Agent
        ↓
เป็น
Alpha Generation Agent
```

ภาพรวม V3 Architecture ที่ผมเสนอ

ผมจะไม่แก้แค่ PPO แต่จะเปลี่ยนเป็น:

```
                Market Data
                     |
        +------------+-------------+
        |                          |
 Technical Features          Macro Features
        |                          |
        +------------+-------------+
                     |
             Market Regime Encoder
                     |
              State Representation
                     |
              Multi-Agent PPO
                     |
        +------------+-------------+
        |            |             |
   Alpha Agent   Risk Agent   Allocation Agent
        |            |             |
        +------------+-------------+
```

```
                |
        Portfolio Controller
                |
         Execution Engine
```

เป้าหมายคือ:

Agent หนึ่งตัวไม่ควรถูกบังคับให้ทำทุกอย่าง

เพราะในตลาดจริง:

- Trader หา alpha
- Risk manager คุม drawdown
- Portfolio manager จัด allocation

มนุษย์ก็ไม่ได้ใช้สมองส่วนเดียวทำทั้งหมด

Phase 1: แก้ Environment ก่อน (สำคัญที่สุด)

ปัญหาปัจจุบันของ V2

Reward:

```
Return
+
Drawdown penalty
+
Cash drag
```

ทำให้เกิด:

```
กลัวแพ้
↓
ถือ cash
↓
Sharpe ดูดี
↓
แต่ไม่มี alpha
```

ดังนั้น Agent ไม่ได้ถูกสอนให้ "ชนะตลาด"

เปลี่ยน Reward Function

ผมเสนอ:

Reward ใหม่

```
Reward(t)

=
Portfolio Excess Return
-
Benchmark Return

+
Risk Adjustment

-
Transaction Cost

-
Drawdown Penalty
```

หรือ:

$$R_t = (r_p - r_b) - \lambda_1 DD - \lambda_2 Turnover + \lambda_3 Alpha$$

ตัวอย่าง

สมมติ:

วันนี้

ตลาด:

```
SPY +1%
```

Agent:

```
Portfolio +1.5%
```

Reward:

```
+0.5%
```

Agent จะเรียนรู้ว่า:

"ไม่ได้แค่ทำกำไร ต้องชนะตลาด"

แต่ถ้า:

ตลาด:

```
-3%
```

Agent:

```
-1%
```

Reward:

```
+2%
```

เพราะมัน protect capital ได้

นี่คือสิ่งที่ hedge fund ทำ

Phase 2: เปลี่ยน Benchmark

ตอนนี้ V2 เทียบ:

```
Equal Weight
```

อย่างเดียว

ผมว่าไม่พอ

เพิ่ม:

Benchmark Layer

Bull market

```
SPY Buy & Hold
```

Growth market

```
QQQ Buy & Hold
```

Defensive

```
60/40 Portfolio
```

Risk parity

```
Volatility Target Portfolio
```

Reward:

Agent ต้องชนะ:

```
SPY
QQQ
60/40
Cash
```

ไม่ใช่แค่ equal weight

Phase 3: เพิ่ม Market Regime Awareness

นี่เป็นจุดที่ผมคิดว่า V2 ขาดที่สุด

ตลาดไม่ได้มี distribution เดียว

มี:

Regime 1: Bull

เช่น:

2020-2021

Characteristics:

```
Low volatility
Positive momentum
Liquidity expansion
```

Strategy:

```
Aggressive allocation
```

Regime 2: Bear

เช่น:

2022

Characteristics:

```
Fed tightening
High yield
Negative breadth
```

Strategy:

```
Cash / defensive
```

Regime 3: Sideway

Strategy:

```
Mean reversion
```

ดังนั้นเพิ่ม:

Regime Encoder

Input:

```
VIX
Yield curve
Fed rate
M2 growth
Credit spread
Market breadth
SPY trend
```

ผ่าน:

```
Transformer Encoder
```

Output:

```
Market state vector
```

เช่น:

```
[0.8 Bull,
0.1 Bear,
0.1 Sideway]
```

ส่งเข้า PPO

Phase 4: เปลี่ยนจาก Single Agent เป็น Multi-Agent

ตอนนี้:

```
PPO

Input
 |
Action:
weights
```

ปัญหา:

มันต้องเรียน:

- timing
- risk
- allocation

พร้อมกัน

ผมจะเปลี่ยนเป็น:

Agent 1: Alpha Agent

หน้าที่:

หา:

```
Which asset will outperform?
```

Action:

```
ranking score
```

เช่น:

```
QQQ +0.8
SPY +0.5
DIA +0.1
Cash -0.5
```

Agent 2: Risk Agent

หน้าที่:

กำหนด:

```
Portfolio exposure
```

Output:

```
0-100%
```

เช่น:

ตลาดปกติ:

```
95%
```

ตลาดวิกฤต:

```
20%
```

Agent 3: Allocation Agent

รวม:

```
Alpha score
+
Risk score
```

เป็น:

```
final weights
```

Phase 5: เพิ่ม Universe

ตอนนี้ V2 มี:

```
SPY
QQQ
DIA
Cash
```

เล็กเกินไป

Alpha เกิดจาก:

"เลือกผิดถูก"

ไม่ใช่:

"ถือสาม ETF"

เพิ่ม:

Equity ETF

```
SPY
QQQ
IWM
EFA
EEM
```

Sector

```
XLK
XLF
XLE
XLV
XLY
XLI
XLP
```

Commodity

```
GLD
SLV
DBC
USO
```

Bonds

```
TLT
IEF
SHY
```

Crypto (optional)

```
BTC ETF
ETH ETF
```

Universe:

30-50 assets

ถึงจะมี alpha opportunity

Phase 6: เพิ่ม Feature Engineering ระดับ Quant

ตอนนี้ feature:

- technical
- macro

ดีแล้ว

แต่ยังขาด:

Momentum

เพิ่ม:

```
12-1 month momentum

6 month momentum

3 month momentum
```

Trend

```
price / SMA200

ADX
```

```
MACD slope

EMA crossover
```

Volatility

```
realized volatility

ATR

volatility regime

volatility ratio
```

Relative Strength

สำคัญมาก

เช่น:

```
QQQ return
-
SPY return
```

Agent ต้องรู้:

"อะไร outperform"

Breadth

เพิ่ม:

```
% stocks above MA50

Advance decline

New highs/lows
```

Phase 7: เปลี่ยน PPO เป็น Algorithm ที่เหมาะกว่า

PPO ดี

แต่ Portfolio optimization เป็น continuous control

ผมจะทดลอง:

Option 1

SAC

ข้อดี:

- sample efficient
- continuous action ดี

Option 2

TD3

เหมาะกับ:

portfolio allocation

Option 3

DreamerV3

อันนี้น่าสนใจที่สุด

เพราะ:

ตลาดเป็น partially observable environment

Dreamer:

เรียน world model

มันสามารถ:

"จำลองอนาคต"

Phase 8: เพิ่ม Memory

ปัจจุบัน:

MLP

คือ:

```
วันนี้เห็นวันนี้
```

แต่ตลาดมี:

```
trend
cycle
regime
```

เปลี่ยน:

```
MLP
```

เป็น:

```
LSTM PPO
```

หรือ

```
Transformer Policy
```

Input:

ย้อนหลัง:

```
60-120 days
```

เช่น:

```
Day -120
...
Day 0
```

Agent จะเห็น momentum เหมือนมนุษย์

Phase 9: เพิ่ม Risk Constraint

แทนที่จะหวังให้ reward คุม DD

ใช้ Constraint:

เช่น:

Maximum drawdown:

```
<15%
```

Volatility:

```
<20%
```

CVaR:

```
<5%
```

ใช้:

Constrained PPO

หรือ:

Lagrangian PPO

Phase 10: Training Strategy

ตอนนี้:

Train

↓

Test

ผมจะเปลี่ยนเป็น:

Curriculum Learning

Stage 1:

ตลาดง่าย

```
2000-2010
```

เรียน:

basic allocation

Stage 2:

เพิ่ม crisis

```
2008
2020
2022
```

Stage 3:

full market

Production Pipeline ใหม่

```
Data Lake
 |
Point-in-Time Dataset
 |
Feature Store
 |
Regime Detector
 |
Multi-Agent RL
 |
Backtest Engine
 |
Stress Testing
 |
Paper Trading
 |
Live Capital
```

Target Metrics ที่ควรตั้ง

ผมจะไม่ตั้ง:

"CAGR สูง"

แต่:

Minimum Gate

Sharpe

```
>1.5
```

Good

```
>2.0
```

Max Drawdown

```
<15%
```

Excellent:

```
<10%
```

Alpha

เหนือ SPY:

```
+5-10% annualized
```

Win Rate

Monthly:

```
>60%
```

Worst Year

ต้อง:

```
ไม่ติดลบหนักกว่า -10%
```

Roadmap ที่ผมแนะนำ

V3.0

แก้:

✅ Reward = Excess Return
✅ Benchmark-relative
✅ Regime features
✅ Bigger universe

เป้าหมาย:

Sharpe:

1.0-1.3

V4.0

เพิ่ม:

✅ LSTM PPO
✅ Multi-agent
✅ Risk constraint

เป้าหมาย:

Sharpe:

1.5+

V5.0 Production

เพิ่ม:

✅ Point-in-time data
✅ Slippage model
✅ Paper trading
✅ Online retraining

เป้าหมาย:

Sharpe:

1.8-2.2

ความเห็นส่วนตัวจากผล V2

ผมคิดว่า V2 ไม่ได้ "ล้มเหลว"

จริง ๆ มันพิสูจน์ว่า:

- Environment ดีขึ้นมาก
- Leakage ถูกแก้
- Risk management ทำงาน

แต่ตอนนี้ Agent ถูกสอนผิด:

มันถูกสอนว่า:

> "อย่าเสียเงิน"

แต่คุณต้องการ:

> "จงหาโอกาสที่ตลาดให้ และชนะมัน"

ดังนั้น V3 ต้องเปลี่ยน Objective จาก:

Capital Preservation RL

เป็น:

Alpha-Seeking Risk-Constrained RL

ซึ่งเป็นแนวทางเดียวกับที่ใกล้กับระบบของ Quant Hedge Fund มากกว่า

ถ้าผมเป็นคนดูแลโปรเจกต์นี้ ผมจะ **ไม่ทิ้ง PPO_V2 แต่ Fork เป็น PPO_V3-A (Alpha Engine)** และเปลี่ยน Reward + Universe + Regime + Memory ก่อนเปลี่ยน Algorithm ครับ.