

深度学习（理论）的**技术科学**道路

旧三论与新三论：从还原、预测到驾驭复杂学习系统

刘方辉

fanghui.liu@sjtu.edu.cn

自然科学研究院/数学科学学院
上海交通大学

CSIAM 深度学习理论专题研讨会 / POSITION TALK

报告结构

定义、道路与问题

01 理论的定义与评价

我们在建设什么样的知识？

02 技术科学的定位与道路

这种理论应当怎样成长？

03 学习系统的具体问题

哪些对象与问题值得研究？

定义知识任务 → 形成理论对象 → 提出可检验的研究问题。

我们讨论的是同一种“理论”吗？

同一句“有没有理论价值”，可能在使用不同标准。

01 严密性

有定理，才算理论？

02 解释力

没有微观机制，
就不算理论？

03 预测力

不能预测，
就没有价值？

04 设计价值

不能立即指导实践，
就不重要？

先明确理论承担什么任务，再讨论怎样评价它。

Scaling law 算不算理论？

一个贯穿全场的例子：预测可以先产生价值，但它提供了什么知识？

经验关系 · 示意

$$\text{loss} \approx L_{\infty} + a \text{ 数据量}^{-\alpha} + b \text{ 参数量}^{-\beta}.$$

由小规模实验外推尚未训练的尺度

条件、参数与外推范围必须明确。

解释微观机制？

未必

预测未观测尺度？

可检验

支持资源配置？

有条件

它形成了什么知识、支持了什么预测，以及这些结论在哪些条件下成立？

可靠的跨尺度预测，可以在机制解释完备之前支持资源决策。

Kaplan et al. (2020); Hoffmann et al. (2022). 公式为给定条件下的示意关系。

理论的取向与功能（数学、规律、计算、设计、决策）

以 scaling law 为线索；六种功能互补，不是价值递增的阶梯。

描述

观察到了什么？

loss 与规模的经验关系

预测

尚未观测的情形会怎样？

外推未训练的尺度

解释

为什么出现这种规律？

指数及修正项从何而来

干预

改变条件后会怎样？

换数据或优化器后的响应

规范 / 设计

为实现目标，应如何选择？

给定预算配置模型与数据

控制

根据反馈，应如何调整？

利用在线观测修改训练日程

有条件地算：系统化、可检验的 scaling law 框架可以成为现象理论/预测性理论。

Kaplan et al. (2020); Hoffmann et al. (2022)。理论功能的区分为本报告框架。

理想的深度学习理论：三要求、四检验

数学上可信、科学上有效、工程上可用：三种互不替代的评价。

数学上可信

假设、推导、证明
与近似误差

模型内结论

科学上有效

观测对应、独立证据
与预测误差

真实系统上的经验主张

工程上可用

目标、约束、代价
与决策后的验证

可靠的决策价值

依主张选择相应检验

可证伪

可预测

可干预

可迁移

不把四项检验作为每篇论文必须同时跨过的门槛。

按所声称的任务评价，并报告适用条件与失效边界。

技术科学：研究一类系统的共性规律

从理论的评价标准，转向围绕哪一类对象建立知识体系。

基础科学	基本规律与数学结构	认识一般结构与基本原理
技术科学	一类系统的共性原理	性能极限、演化规律与设计原则
工程技术	具体目标与资源约束	构建、运行和改进一个系统

相互作用、边界交错；不是价值高低的排序。

技术科学的**核心**是对问题的认识以及建模。

技术科学有自己的基本问题，不是“数学应用”或“工程经验”的别名。

钱学森，《论技术科学》（1957）；三层框架为意译与提炼。

深度学习的技术科学道路

本报告的立场：连接不同理论贡献，而不是用单一标准筛选它们。

**以学习系统的共性问题为对象，
形成可积累、可检验的知识框架。**

基本认识

结构、机制与能力极限

经验规律

预测、适用范围与失效边界

设计原则

目标、约束、权衡与可靠性

不是每篇论文包办一切，而是让不同贡献能够相互连接。

本报告主张：发展深度学习的技术科学理论。

深度学习既是研究对象，也是设计对象

不仅研究系统如何运行，也研究它可以被设计成什么。

研究对象

解释机制 · 刻画规律

识别能力与性能极限

设计对象

改变结构与训练方式

研究目标、约束与性能权衡

模型 × 数据 × 优化器 × 算力 × 环境

深度学习的技术科学理论

研究深度学习系统的共性原理、演化规律与设计原则。

认识问题、建立模型、分析计算与实验、提炼共性规律，并把这种理解转化为可检验的设计判断。

Simon, The Sciences of the Artificial (1996); 本报告的研究定位。

旧三论：把具体问题变成理论对象

信息论、控制论、系统论提供不同的抽象方式，而不是现成答案。

信息论

信道能传多少信息？

熵 · 容量 · 编码

把通信问题变成
可以比较的量

控制论

怎样在扰动下实现目标？

反馈 · 状态 · 稳定性

把动态行为变成
分析与设计问题

系统论

部分如何形成整体？

关系 · 层次 · 耦合

把相互作用与
整体行为作为对象

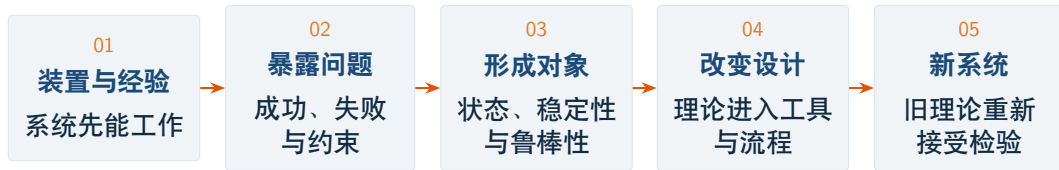
共同点在于抽象方法，并非三者拥有相同的工程起源。

借鉴的是构造理论对象的方法，哪些差异可以忽略，哪些关系可以保留，最后希望回答什么问题。

Shannon (1948); Wiener (1948)。三者不具有完全相同的历史起源。

控制理论是在工程循环中形成的

先看一条贯穿控制史的线索：实践与理论相互推动，而非单向接力。



历史线索：应用、失败、抽象与新设计不断相互推动。

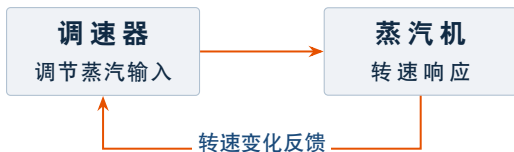
实践提出问题，理论形成结构，新的设计再带来新的问题。

IEEE Control Systems Society, History; 研究循环为本报告提炼。

从调速器失稳，到稳定性问题

机械时代 | 1788 → 1868：从反馈装置到可分析的稳定条件。

1788 瓦特离心调速器



1868 Maxwell · On Governors

把振荡转化为可分析的稳定性问题

微分方程 → 参数条件 → 稳定性

(1892 Lyapunov 稳定性提出，但并未引起注意)

经验调节：为什么调节得更强，系统反而会振荡？

可重复的成功与失败，都能成为提出共性理论问题的起点。

IEEE CSS, History: Watt (1788); Maxwell, On Governors (1868)。

理论首先改变什么？解释失败、划定边界

从稳定性分析走向系统设计，不是一次完成的跨越。

分析

给定装置与参数

判断为什么振荡、何时稳定

综合 / 设计

给定目标与约束

选择结构、参数与反馈规律

01 解释失败

为什么会出问题？

02 划定边界

哪些参数不合适？

03 缩小试错

Routh-Hurwitz 判据

下一次怎样更有依据？

从“可分析”走向“可设计”，需要进一步把理论组织为设计方法。

理论未必直接发明装置，也可以通过压缩试错空间改变实践。

Maxwell (1868); 钱学森 (1957)。分析与设计的区分为本报告提炼。

经典控制：找到可测量、可设计的有效量

电气时代 | 1920s–1940s：通信与伺服系统推动频域方法的发展。



Black：负反馈放大器

Nyquist：稳定性分析

Bode：频域设计

研究期待：从更多“Maxwell 式分析”，走向“Nyquist/Bode 式设计语言”。

找到连接测量、分析与设计的有效描述，是技术科学的重要任务。

Black, IEEE oral history (1977); Nyquist、Bode 的频域工作。

现代控制：新问题需要新的描述语言

航天与计算机时代 | 1950s–1960s：状态、观测与最优决策。

线性离散系统（示意）

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

$$y_t = Cx_t + v_t.$$

x : 状态 u : 输入 y : 观测

w/v : 过程扰动 / 观测噪声

应用仍需：非线性扩展与数值实现

01 状态

系统内部如何演化？

02 可观性与估计

由观测可以推断什么？

03 可控性

输入可以改变哪些状态？

04 最优决策

如何权衡状态误差与控制代价？

范式变化可能来自新的状态表示，而不只是把旧工具用得更强。

Kalman (1960); McGee & Schmidt, NASA-TM-86847 (1985)。

鲁棒性：模型内最优，不等于失配下可靠

1970s–1980s：模型误差成为理论必须正面处理的对象。

标称模型

已知动力学与目标函数

最优策略 → 模型内保证

实际系统

未建模动态、时延与实现误差

允许哪些失配？保证还成立吗？

Doyle · 1978

一般 LQG 设计没有统一的非零增益 / 相位裕度保证。

不是“LQG 都不稳定”，而是最优性本身不能替代鲁棒性论证。

不是要求模型完全正确，而是明确误差结构与保证能够保持的范围。

Doyle, IEEE TAC (1978)

MPC：实践与理论不是单向接力

工业约束推动方法，也持续推动可行性、稳定性与鲁棒性研究。



DMC 等工业应用



可行性、稳定性与鲁棒性分析

实践价值不必等待理论完备；可靠性保证仍须说明条件与边界。

Rawlings, Mayne & Diehl (2017), §2.11; DMC: Cutler & Ramaker (1980)。

理论成熟后，如何保持问题来源？

工具的成功，与研究议程的更新，是两件不同的事。

一个反思问题

**Control theory
is dead?**

— 哈佛大学 何毓琦教授

一个反思问题，
不是历史定论。

工程的成功

成熟方法进入软件、仿真
与工程流程。

研究的追问

新的系统、约束与失败，
是否正在进入理论问题？

需要警惕的不是数学，而是理论与问题来源脱节。

Murray (ed.), Control in an Information Rich World (2002)。提问不作为直接引语。

深度学习理论应当走什么道路？

从控制史得到的不是现成答案，而是建立理论的方式。



下一部分：学习系统的状态、输入、观测与有效描述是什么？

以数学建立结构，以实验检验边界，再以预测与设计推动实践。

钱学森 (1957); Murray (ed., 2002)。研究道路为本报告主张。

loss 和梯度，是观测还是足够的状态？

关键不是测到了多少量，而是保留了多少预测所需的信息。

读数 / OBSERVATION

$$y = h(s)$$

loss、梯度、谱统计

候选描述 / EFFECTIVE STATE

$$z = \varphi(s)$$

保留预测所需信息

相近的候选状态 + 相同的未来输入

能否支持相近的未来响应预测？

先限定目标、预测时域与输入范围；随机训练比较响应分布或预测误差。

有效状态不是“与结果相关的指标”，而是面向预测目标的信息概括。

新三论：提出问题，不预设答案

耗散结构、协同论与突变论提供研究问题，而不是现成解释。

耗散结构

稳定模式如何维持？

哪些外部驱动与耗散
维持训练中的稳定模式？

协同论

有效变量如何出现？

是否有时间尺度分离？
少数变量能否组织演化？

突变论

定性变化如何发生？

EoS、grokking 等现象
是否存在可预测的边界？

本页是问题生成器：不把数据输入直接等同于热力学通量，也不把突变直接等同于相变。

类比的价值在于提出假设；理论的价值还需要定量检验。

深度学习是否存在真正的序参量？

在协同论启发下提出检验要求，而不是把相关性改名为序参量。

结构假说

候选慢变量

特定阶段或临界附近



快速变量松弛

检验时间尺度分离



有效动力学

给出近似与误差

需要通过的检验

01 近似闭合

保留的信息能否预测未来演化？

02 干预响应

改变输入后，关系是否仍有效？

03 适用范围

描述何时出现，何时开始失效？

这是对候选有效变量的研究要求，不是“序参量”的跨学科统一定义。

“序参量”应当是一项有结构条件的理论主张，而不只是一个好用的指标。

复杂系统不保证存在低维闭合

寻找结构条件，也允许谱、分布或带记忆的有效描述。

可能的约简来源

时间尺度分离

快变量松弛，保留慢变量

高维集中

自平均与稳定统计量

结构与目标

对称性、冗余与目标投影

需要警惕的遗漏

历史依赖与滞后 / 亚稳态与稀有事件 / 方向耦合 / 阶段与分布变化

低维无记忆描述不足

→ 更多变量、谱 / 分布状态，或带记忆的响应关系

目标不是先把维数压低，而是保留预测所需的结构，并控制遗漏误差。

有效状态可能随训练阶段变化

以下是待验证的情景，不是一般神经网络必然经历的四个阶段。

初期：高维瞬态

多方向同时变化
未必有固定低维概括

有效方向形成

表征、margin、有效秩
可能成为候选变量

临界附近

少数边缘稳定方向
可能主导有限时间响应

表征重组

原有变量或响应关系
可能需要更新

要检验的不只是阶段内拟合，还有阶段边界、模型切换与预测误差。

不预设全程低维；识别有效描述何时出现、何时切换、何时失效。

“相变”与“涌现”需要定量判据

先明确要检验的主张，再选择变量和实验。

形状相似还不够

“曲线突然改变”

“不同实验看起来相似”

这些是现象线索，
还不是机制判据。

01 变量

什么量变化，为什么选择它？

02 参数与边界

改变什么触发变化，边界怎样移动？

03 事前预测

留出条件下能预测什么？

04 替代解释

有限尺度、阈值或噪声能否解释？

类比可以启发理论，但不能替代变量、方程、预测与失效条件。

怎样检验一种有效描述？（控制变量法）

从已有轨迹上的吻合，走向留出条件下的响应预测。



技术归零

定位准确、机理清楚、问题复现、措施有效、举一反三

有价值的预测失败，应帮助定位遗漏的信息、近似误差或适用条件。

把研究道路落实为四类具体问题

每个方向都要明确：有效变量是什么，哪些预测可能失败。

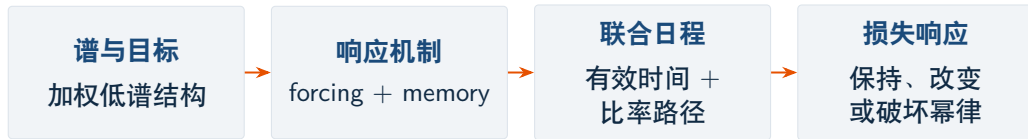
研究问题	候选变量 / 结构	事前检验
Scaling laws	指数、修正项与记忆结构	新尺度或新日程下，保持什么？
μ P transfer	参数化与跨宽度不变量	迁移后哪些量保持，条件何时失效？
WSD / 训练时长	有效时间与日程历史	改变步数或调度，响应如何变化？
EoS / Muon / batch	临界模态与稳定边界	改变学习率或批大小，边界如何移动？

每一行都是研究问题，不是已经完成的统一理论。

不是增加四条孤立研究线，而是让不同问题共享可检验的理论要求。

一个实例：3 + 3 + (2) Scaling Laws for LLM Pre-training

Yichen Wang, Fanghui Liu, Yudong Chen. *From Spectra to Joint Schedules in LLM Pre-training: 3+3(+2) Scaling-Law Regimes*.
[arXiv:2609.40148](https://arxiv.org/abs/2609.40148).



有效时间

$$T_t = \sum_{s < t} \eta_s$$

学习率累积：优化进展

比率路径

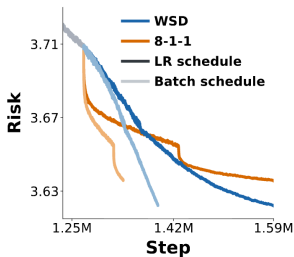
$$r_t = \frac{B_t}{\eta_t}$$

其倒数控制单位有效时间的噪声注入

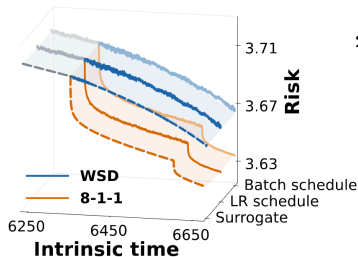
适用起点：冻结随机特征 + noisy online SGD; (T, r) 不是无记忆的完整训练状态。

可解代理提出响应坐标；其价值还要由复杂系统上的预测检验。

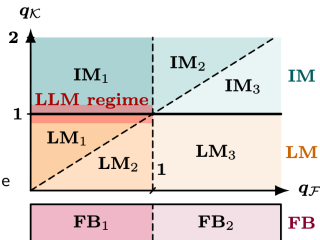
一个实例: $3 + 3 + (2)$ Scaling law regimes and joint schedules



(a) Loss vs. step



(b) Loss vs. intrinsic time

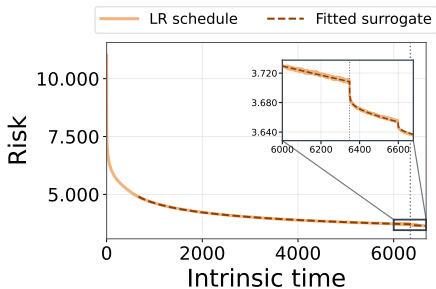
(c) $3 + 3(+2)$ map

有效描述验证；右侧分区图仅作理论结果概览。

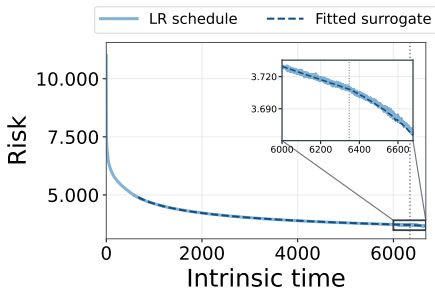
loss 的有效描述。能否预测？

一个实例: $3 + 3 + (2)$ Scaling laws for LLM Pre-training: Prediction

Motivated by Li et al. *Functional scaling laws*, NeurIPS '25.



(a) 8-1-1 fit



(b) WSD prediction

8-1-1 拟合 $\rightarrow$ 固定参数 $\rightarrow$ 留出 WSD 预测。

固定模型与参数以后，仍能预测什么？这比逐条重新拟合更强。

从预测有效，到可靠设计，还差什么？

回扣第一部分：预测、解释、干预与反馈控制承担不同任务。

识别与监测

有效状态与不确定性

怎样发现模型开始失效？

目标与约束

loss、能力、成本

按 token、算力还是时间比较？

决策与反馈

何时调整，幅度多大

误差与扰动下是否仍可靠？

下一步的检验对象：在声明的目标、预算与约束下，决策之后的真实结果。

预测成功可以成为设计的起点，但不能替代目标、约束与闭环检验。

总结：深度学习的**技术科学**道路

微观、中观/介观与宏观是互补的描述尺度；技术科学连接问题、认识与设计。

核心：认识问题与建模

从共性问题出发，构造有效描述。

明确对象、关键变量与需要保留的结构。

途径：分析、计算与实验

相互启发、检验与修正。

推导关系，提出新预测，识别模型误差。

积累：共性规律与适用边界

解释机制，刻画极限与响应规律。

哪些关系能够保持，何时开始失效？

目标：增进认识，支持设计

形成可检验的预测与设计判断。

例如：在同等预算下选择并验证训练日程。

需要警惕的不是数学，而是理论与问题来源及经验检验脱节。

以问题牵引建模，以分析、计算与实验提炼规律，以预测与设计返回实践。

预训练尺度律的形成机理与调控

三个相互关联的问题：有效状态、规律迁移与干预预测。

- 01 有效状态 哪些观测组合足以支持训练响应预测？
- 02 规律迁移 哪些规律能跨尺度、数据与训练阶段保持？
- 03 干预预测 能否事前预测改变条件的响应及失效边界？

预训练尺度律的形成机理与调控

三个相互关联的问题：有效状态、规律迁移与干预预测。

- 01 有效状态 哪些观测组合足以支持训练响应预测？
- 02 规律迁移 哪些规律能跨尺度、数据与训练阶段保持？
- 03 干预预测 能否事前预测改变条件的响应及失效边界？

Q & A

fanghui.liu@sjtu.edu.cn

www.lfhsgre.org

什么是理论？一个工作性定义

本报告采用的工作性定义，不作为唯一的哲学定义。

理论

围绕一类对象或问题，由明确的概念、假设与关系组织起来，能够系统地产生结论的知识框架。

对象与概念



假设与关系



可推导结论

若声称描述真实系统，还需说明

如何对应观测 / 怎样接受检验 / 适用范围在哪里

单条定理或拟合曲线可以是理论的一部分，但不自动构成完整理论。

理论的类型：不同的研究取向

区分研究取向，是为了辨认其组织依据与论证方式。

理论取向	主要组织依据	典型例子
数学 / 形式	公理、假设与逻辑推导	泛化界、优化收敛定理
经验 / 科学	观测、实验与可检验预测	scaling 规律、经验动力学模型
计算	可计算性与资源复杂度	计算复杂性、PAC 学习
工程 / 设计	目标、约束与可靠性要求	控制理论、编码理论
规范	目标、偏好与决策准则	决策理论、最优控制

同一理论可以同时依靠数学推导、经验判断和设计论证。

这些不是互斥类别；同一套理论可以同时具有多种取向。