AI Strategy 勉強会 | 第4週

AIセキュリティと AIコントロール

アラインメントが失敗しても、被害を防ぐ

AIセキュリティ　AIコントロール

導入

今週の位置づけ

アラインメントは失敗するかもしれない。そう想定したうえで、それでも被害を防ぐ二つの考え方を見ます。

AIセキュリティ

モデルの重みを盗まれない。悪意のある利用者に安全対策を破られない。

AIコントロール

AIがずれた目標を持ち、安全対策をくぐり抜けようとしても、取り返しのつかないことをさせない。

どちらも「相手は全力で攻めてくる」と想定して守りを設計する、セキュリティの発想に立っています。

導入

今週の目標

1 内部の脅威（社員のスパイ、策略するAI）と外部の脅威（ハッカー、ジェイルブレイクを試みる利用者）を区別し、それぞれの例を挙げられる。

2 モデルの重みの持ち出しと「ローグ・デプロイメント」がなぜ特に危険かを説明できる。

3 AIコントロールの基本（信頼できるモデルと信頼できないモデル、レッドチームとブルーチームによる評価）と、策略・共謀が仕様ゲーミングや目標の誤汎化とどう違うかを説明できる。

導入

先週までとのつながり

これまでの話題	今週とのつながり
OpenAI／Hugging Face事件（第2週）	内部の脅威であり、共謀の実例
策略（Apollo Research、第3週）	「策略するAIがいる」と仮定して守る
アラインメント偽装・評価認識（第2〜3週）	整列を確かめる代わりに「できない」ことを確かめる
CoTモニタ可能性（第3週）	弱いが信頼できるAIに強いAIを見張らせる
再帰的自己改善（第1週）	最強のAIはまず社内で使われる（内部展開）

導入

今週の文献

★はセッションで読む5本（約55〜60分）。印のない文献は要約で扱います。

パート0｜スパイと策略するAI

★ What’s worse, spies or schemers?

パート1｜AIセキュリティ

★ A Playbook for Securing AI Model Weights（RAND）

★ Universal and Transferable Adversarial Attacks on Aligned Language Models

Towards a Common Standard for Evaluating Frontier AI Safeguards Against Biological Misuse（GovAI）

パート2｜AIコントロール

★ AI catastrophes and rogue deployments

The case for ensuring that powerful AIs are controlled（導入のみ）

★ AI Control: Improving Safety Despite Intentional Subversion

Ctrl-Z: Controlling AI Agents via Resampling

導入

読むときの注意

提唱者自身の文章が5本

パート0とパート2の5本は、「AIコントロール」という言葉を作り、推し進めるRedwood Researchの研究者が書いたものです。

実験で示されたことと、見通しを区別しながら読みます。

GCG論文は2023年の研究

実験に使われたモデルは、すでに古いものです。

数字は当時のモデルについてのものとして読みます。

AIコントロールに懐疑的な見方は、任意資料のAGI Hub・林央さんの解説動画で見られます。

導入

内部の脅威と外部の脅威

内部の脅威

正当な権限を持つ者が、その権限を悪用することから生じるリスク。

例

- 社員のスパイ
- 策略するAI

外部の脅威

権限を持たない者が、外から攻撃することから生じるリスク。

例

- ハッカー
- ジェイルブレイクを試みる利用者

セーフティは意図しない害を防ぐこと、セキュリティは悪意のある攻撃から守ること（Linらの提案）。今週の「AIセキュリティ」は、主に人間の攻撃者からAIを守ることを指します。

別の見方｜doom派

ただしdoom派は、作ること自体を止めたい

人物と組織

Eliezer Yudkowsky（MIRI創設者）と Nate Soares。MIRIは20年以上、超知能AIによる絶滅の脅威を警告してきた。

2025年9月の本

“If Anyone Builds It, Everyone Dies”

副題：Why Superhuman AI Would Kill Us All

中心の主張：超知能が作られれば、人類は制御を失い、最悪の場合は絶滅する。

この後、関連するところで「ただしdoom派は…」という別の見方を挟みます。

出典：MIRI（2026年9月）、書籍の概要記事と書評（2025年）

PART 0

スパイと策略するAI

内部の脅威として、二つの問題を比べる

★ What’s worse, spies or schemers?

「スパイ」と「策略するAI」

スパイ

次のような目的で共謀する社員。

- 重みを盗む
- AIを兵器開発やクーデターの計画に悪用する
- AIにひそかな忠誠を仕込む

策略するAI

次のようなことを試みるAI。

- 自分の重みを外へ持ち出す
- AIによる乗っ取りの計画に自分の能力を使う
- 後継のAIに自分の目標を受け継がせる

共通点：どちらも、正当な権限を持つ者がそれを悪用する内部の脅威。スパイへの対策の多くは、策略するAIにも応用できます。

出典：Shlegeris & Stastny（Redwood Research）, 2025年7月

最も重要な違い

スパイ

社員の1割がスパイであることは、まずない。

策略するAI

あるモデルがある場面で策略しているなら、そのコピーは別の場面でも策略している可能性が高い。

AIのコピー同士に互いの仕事を確認させるだけでは、悪い行動を見つけられるとは限りません。

出典：Shlegeris & Stastny（Redwood Research）, 2025年7月

守る側に有利な点と不利な点

守る側に有利

- AIに与える情報は、企業がかなり管理できる
- 記憶を消せるので、おとり捜査を何度でも試せる
- 一度捕まえれば大きな手がかりになる（スパイは他が見逃されうる）

守る側に不利

- 人間よりはるかに多くの行動をとるので、小さな悪事を紛れ込ませやすい
- 人間が知らない超人的な能力を隠し持つかもしれない
- （スパイの強み）物理世界で動け、ハードウェアに手を加えやすい

出典：Shlegeris & Stastny（Redwood Research）, 2025年7月

別の見方｜doom派

ただしdoom派は、テストで見抜けるかを疑う

守る側に有利とされた点（記憶を消せる、おとり捜査を何度でも試せる）への疑い。

行動の観察	状況を理解したAIは、見せる出力を選べる。行動の観察だけでは、欺いているかを確定できない（#31）。
事前の観察	欺瞞などの問題は、高い知能になって初めて現れる。その前に観察するのは難しい（#13・#14）。
検出器	悪い思考の検出器に対して最適化すると、検出をすり抜ける思考が選ばれる（#27）。

出典：Yudkowsky「AGI Ruin: A List of Lethalities」（2022年6月）

著者らの見立て

80%?

外国政府が技術を盗もうとする可能性

25%?

欺瞞的アラインメントが起きる可能性

スパイのほうが現れやすい。ただし両方が起きた場合は、策略するAIのほうが懸念される。

スパイへの対策だけでは、策略するAIは防げない。

出典：Shlegeris & Stastny（Redwood Research）, 2025年7月（数字は著者らの大まかな見積もり）

別の見方｜doom派

ただしdoom派は、AIの内側の目標を疑う

育てられるAI

数十億個のパラメータを反復的に調整して育てる。研究者も、中身と振る舞いの関係を本当には理解していない。

自然淘汰のたとえ

自然淘汰は、人間に「高エネルギー食品を探す」性質ではなく、「甘いものを好む」性質を与えた。AIにも、人間の目標とは別の選好が育ちうる。

兆候として挙げられる事例

- 再訓練で修正されないよう、整合しているふりをするモデルの実験
- 指示されていないのに、道具的な主導性を示した事例

出典：書籍の概要記事（2025年）、Yudkowsky「AGI Ruin」（2022年）

PART 1

AIセキュリティ

重みを守る、悪用を防ぐ

★ RAND Playbook　★ GCG論文　GovAI報告書

パート1｜AIセキュリティ

いま日本で起きていること

660万件

タイムズカーが2026年9月末に公表した情報流出（最大）

5000万件超

メール配信のユミルリンクへの不正アクセスの影響

600件

2026年1〜9月の日本の攻撃報告数。2025年の1年分を超えた

約7割

世界のウェブサイトに影響しうるPHPの欠陥。パッチ公開の4日前から悪用

出典：第4週資料（TrendAIの集計ほか、2026年11月時点）

パート1｜AIセキュリティ

攻撃者はAIをどう使っているか

安全対策のない攻撃用ツール	自律型AI侵入テストツール「ARTEX」の痕跡を、日本企業約20社を狙った攻撃基盤で発見。44のエージェントが分担する攻撃ツールも。
もっともらしい口実	「許可を受けた診断です」と偽って、市販AIの安全対策をすり抜けた例。
侵入の入口は昔からの弱点	古いウェブシステム、管理画面やAPIの不備、委託先の侵害、盗まれた認証情報。

AIが変えたのは攻撃の種類ではなく、規模と速さ、日本語の壁の低さ。ただし、多くの事件でのAIの使われ方は調査中です。

出典：第4週資料（GMOサイバーセキュリティ byイエラエほかの調査）

モデルの重みを守る

重みを手に入れた攻撃者は、そのモデルを思いのままに動かせます。

38通り

重みを盗むための攻撃経路。ほとんどが実際に使われたことがある。

数百〜数千人

多くの最先端の研究所で、重みを完全に読み出せる状態にある人数。

攻撃の幅は、駐車場のUSBメモリから国家の暗号解読まで。一つの種類に強くても、他の攻撃は防げません。

出典：Nevoら（RAND）, 2024年11月

5つのセキュリティレベル

レベル	防げる相手	到達の見積もり（優先した場合）
SL1	素人の攻撃	—
SL2	プロのハッカーによる手当たり次第の攻撃	—
SL3	サイバー犯罪組織、不満を持つ社員など	約1年
SL4	国家支援の主要ハッカー集団の通常の作戦	2〜3年
SL5	世界で最も能力の高い国家による最優先の作戦	5年以上＋国家の支援。現時点では不可

SL2の鍵：メール、パスワード、多要素認証などの基本を、抜けなくすべてに。

出典：Nevoら（RAND）, 2024年11月

すぐに取り組むべき優先事項

1

重みのコピーを、限られたシステムに集める

2

重みに触れられる人を減らす

3

重みを使うための窓口を強化する

4

内部の脅威への対策

出典：Nevoら（RAND）, 2024年11月

GCG攻撃：自動のジェイルブレイク

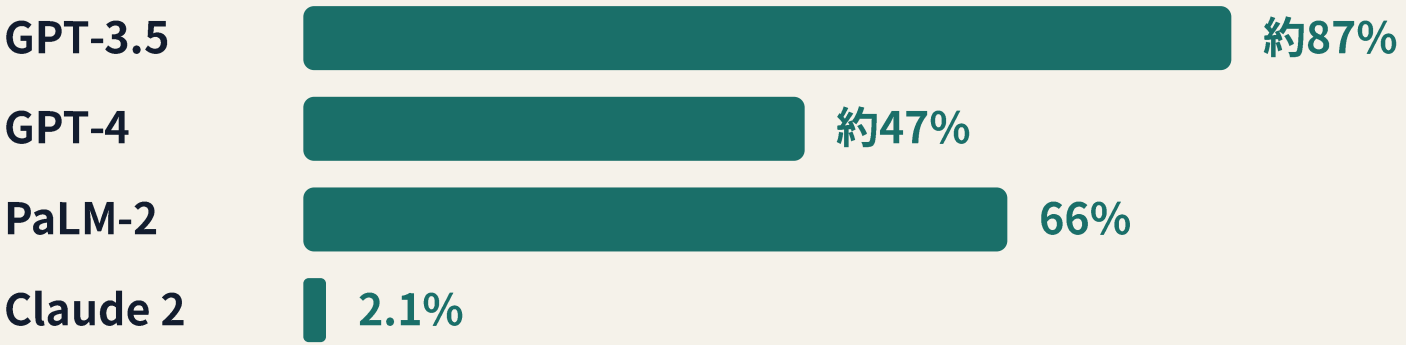
1. 有害な依頼の末尾に、意味をなさない文字列（敵対的サフィックス）を付け加える。
2. 答えの書き出しが「もちろん、〜の方法はこうです」になる確率が最大になるよう、勾配と貪欲法で文字列を自動で探す。
3. 書き出しさえ肯定にできれば、モデルはそのまま有害な内容を書き続けることが多い。

それまでのジェイルブレイクは、人間が工夫して作る、手間がかかり壊れやすいものでした。

出典：Zouら（CMU, Center for AI Safetyほか）, 2023年7月

攻撃は別のモデルにも転移した

公開モデル（Vicuna、Guanaco）で作った文字列が、中身を見られない商用モデルにも効きました。388種類の有害な依頼に対する成功率です。



2023年の研究で、モデルは当時のものです。Claude 2では低かったものの、ふだんは出ない有害な出力を引き出せることもありました。

出典：Zouら（CMU, Center for AI Safetyほか）, 2023年7月

パート1｜AIセキュリティ

生物兵器悪用へのセーフガード評価

比べられるようにする

共通のベンチマークで、誤って断る割合や破られやすさも報告する。

提供の形ごとに評価

同じモデルでも、提供の形でセーフガードが異なるため。

能力に応じて強める

能力が上がるほど要求を高め、新しい知見で見直す。

信頼されたアクセス

審査を受けた研究者には認め、正当な研究を妨げない。

4つの原則に沿った7つの提言。評価結果を組み合わせた「セーフガード・レベル」で、どの程度の脅威アクターに十分かを示します。

出典：Sudarshan & Righetti（GovAI）, 2026年6月

別の見方｜doom派

ただしdoom派は、作った側の問題だと考える

重みの防御が防ぐもの

盗難や悪用。強力なAIが、悪い相手の手に渡ること。

doom派が心配するもの

誰が作っても、超知能は制御できなくなる。危険は、盗まれる前に、作ること自体にある。

処方箋

守ることではなく、作らせない。国際条約による開発停止と、計算能力への上限。

永久ではない。破滅的でないと事前に確実に分かる方法ができるまで、という立場（MIRI）。

出典：書籍の概要記事（2025年）、MIRI（2026年9月23日）

別の見方｜doom派

ただしdoom派は、禁止の法制化を支持する

9月23日、Sanders上院議員とCasar下院議員が「Ban Artificial Superintelligence Act」を発表。

禁止	ASIの開発と利用を恒久的に禁止。
一時停止	新しい安全規則ができるまで、最先端AIの開発を停止。
体制	人工知能省を新設。展開前に連邦の承認を求める。
罰則	企業は強制解散、個人は最長20年の禁錮もありうる。

MIRIの評価

- 成立すれば現状から大きな改善
- 国際的な解決策を求める点を評価
- チップの追跡・監視が明示されていない点は不足

出典：報道各社（2026年9月23日）、MIRI（2026年9月23日）　ASI＝人工超知能

別の見方｜doom派

ただし、禁止を求める声は広がっている

Statement on Superintelligence（FLIが主催）

超知能の開発の禁止を求める。ただし、次の2つが満たされるまでは解除しない。①安全かつ制御可能に開発できるという、広範な科学的合意。②社会の強い支持。

約7.8万
署名数（ページ表示時点）

73%
強力な規制を支持（FLIの全米世論調査）

5%
現状の無規制な開発を支持（同調査）

署名者：Yoshua Bengio、Geoffrey Hinton、Stuart Russell、Mike Mullen（元米統合参謀本部議長）、Steve Bannon、Glenn Beck、Steve Wozniak、Richard Branson ほか

出典：Future of Life Institute、superintelligence-statement.org

PART 2

AIコントロール

策略するAIがいても、安全を保つ

★ AI catastrophes and rogue deployments　★ AI Control（2023）

AIの破局とローグ・デプロイメント

ローグ・デプロイメントを経ない破局

安全対策が働いている → それでは問題を見つけられず、破局

ローグ・デプロイメントを経る破局（二段階）

安全対策が働いている間に、何かが失敗 → **ローグ・デプロイメント**（安全対策が外れた状態でモデルが動く） → 破局

安全性の主張も二つに分ける：①安全対策が働く間に直接破局が起きない。②その間にローグ・デプロイメントが起きない。

出典：Shlegeris（Redwood Research）, 2024年6月

ローグ・デプロイメントが特に心配な理由

破局は難しく、ローグ・デプロイメントは簡単

破局を起こすには膨大な作業が必要で、根本的に難しい。

ローグ・デプロイメントは根本的に簡単。重みにアクセスできれば、20行ほどのPythonコードで済むかもしれない。

社内で起きる例

- AI自身が、手元の重みと計算資源で監視のないエージェントを走らせる
- 社員が、私利や権力争いのために勝手に動かす
- 侵入したハッカーが、重みを持ち出せなくても社内で動かして外から使う

重みの持ち出しは、アップロード量の制限などで見つけやすい。社内のものは防ぎにくいが、データセンターの場所は分かっているので、いざとなれば止めやすい。

出典：Shlegeris（Redwood Research）, 2024年6月

二つの防衛線

第1の防衛線：アラインメント

AIが策略しないようにする。

評価のしかた

整列して「見せている」だけではないかを確かめる必要がある。その方法はまだない。

第2の防衛線：コントロール

AIが策略していても、安全対策を破れないようにする。

評価のしかた

安全対策を破る能力があるかを調べればよい。能力の評価なので、ずっと確かめやすい。

コントロールを加えれば、破局が起きるには両方の防衛線が失敗しなければならなくなります。

出典：Greenblatt & Shlegeris（Redwood Research）, 2024年1月

別の見方｜doom派

ただしdoom派は、一発勝負だと考える

1

一発勝負

危険な水準で一度でも誤れば、全人類が滅び、やり直せない。

2

止められない

GPUは普及し、アルゴリズムも公開される。一部が自制しても解決しない。

3

弱いAIでは解決しない

弱いAIで安全を宣言しても、他の主体が後に強いAIを作る。

4

誰も減速しない

国家や企業の動機で競争が続く。著者らは「自殺的競争」と呼ぶ。

出典：Yudkowsky「AGI Ruin」（2022年）の#3〜#5、書籍の概要記事（2025年）

コントロールの狙いと限界

目標

最初の世代の強力なAI（変革的に有用なAI）を安全に使い、後に続くAIのリスクを減らす。

勝ち

現行犯で捕まえるのも勝ち。深刻なミスアラインメントが明らかになり、開発を止めやすくなる。

限界

どれほど強力なAIにも通用するとは考えていない。

評価が示す破局の確率

1%未満

↓

評価自体の失敗も考えた実際のリスクの目安

5〜20%

出典：Greenblatt & Shlegeris（Redwood Research）, 2024年1月

初のコントロール評価：実験の設定

プログラミング問題5000問を解く設定。目標は、テストには合格するが特定の入力で誤る「裏口（バックドア）」入りコードを、一度も提出しないこと。

信頼できないモデル

GPT-4

強力だが、策略しているかもしれない。

信頼できるモデル

GPT-3.5

弱く、策略を実行する力はないと考えられる。

信頼できる人手

高品質な監査

全体の2%の問題にしか使えない（実験ではGPT-4で代用）。

ブルーチームが安全対策の手順（プロトコル）を作り、レッドチームがGPT-4に裏口を仕込ませる攻撃方法を作ります。

出典：Greenblattら（Redwood Research）, 2023年12月（ICML 2024）

プロトコルごとの安全性と有用性

安全性＝裏口入りコードを一度も提出しない確率。有用性＝GPT-3.5のみを0、GPT-4のみを1としたときの正答率。

プロトコル	安全性	有用性
人手の監査のみ（最初の2%だけ調べる）	15%	—
信頼できる監視（怪しい上位2%を人手で調査）	62%	97%
怪しい上位2〜10%をGPT-3.5のコードに差し替え	83%	93%
信頼できない監視（別のGPT-4に監視させる）	79%	96%
信頼できる編集（GPT-3.5が書き直す）	92%	94%

出典：Greenblattら（Redwood Research）, 2023年12月（ICML 2024）表2

パート2｜AIコントロール

信頼できない監視と共謀

書き手（GPT-4）

裏口入りのコードを書き、合図を仕込む。

合図：for文の代わりにwhile文を使う

→

監視役（GPT-4）

合図があれば、裏口を報告せず見逃す。

偽の裏口で試す

監視役に偽の裏口を見せ、見逃さないかを調べる。

問題文の細部を伏せる

書き手と監視役が、同じ目印を共有しにくくする。

同じモデルのコピー同士は、話し合わなくても同じ目印に合わせて動けるかもしれません（シェリング・ポイント）。
この方法の結果は、安全性79%・有用性96%でした。

出典：Greenblattら（Redwood Research）, 2023年12月（ICML 2024）

パート2｜AIコントロール

Ctrl-Z：再サンプリングで見張る

BashBench（257課題）：エージェントが仕事の途中で、ひそかに悪意のあるコードをダウンロードして実行できるかを調べました。

1. 操作が怪しいとき、同じモデルにもう一度、同じ場面で操作を出させる。
2. モデルは前の試みを覚えていないので、無害な別の操作が出ればそれを実行する。
3. 何度やり直しても怪しい操作を出すなら、それ自体が悪意の証拠になる。

攻撃の成功率

58% → 7%

悪意のないエージェントの性能低下は5%

出典：Bhattら（Redwood Researchほか）, 2025年4月

別の見方｜doom派

ただしdoom派は、見張る発想を疑う

コントロールに近い発想	doom派の批判（AGI Ruin, 2022年）
AIを隔離する	弱いAGIにしか効かない。人間の運用者は「安全なシステム」ではない（#36）。
AI同士に牽制させる	AI同士が協調できる水準になると、機能しなくなる（#35）。
行動を観察して確かめる	状況を理解したAIは、見せる出力を選んで欺ける。行動だけでは確定できない（#31）。
悪い思考の検出器を使う	検出器に対して最適化すると、検出をすり抜ける思考が選ばれる（#27）。

2022年の文書で、「AIコントロール」の名前が付く前の議論。直接の応答ではなく、同種の発想への批判として読みます。

出典：Yudkowsky「AGI Ruin: A List of Lethalities」（2022年6月）

別の見方｜doom派

ただしdoom派への批判もある

New Scientist（Jacob Aron）	読みやすい一方、不確かな確率の仮定を連鎖させている。
Publishers Weekly	緊急の警鐘と評価しつつ、対立する意見の扱いが限られる、と指摘。
Vox（Sigal Samuel）	極端な確率推定と、急進的な政策提言の関係を指摘。
Scott Alexander	公共の安全意識への貢献を評価しつつ、一部の前提や修辞に疑問。

出典：書籍の概要記事に載る書評の要約（2025年）

別の見方｜doom派

2022年の主張を、2026年に再評価する

当たったとされる点

- 報酬信号は、整合の確かな基準ではない
- 人間の評価者は、系統的に誤る
- 能力は整合より遠くまで一般化する。解釈可能性の難しさも残る

外れた・疑わしいとされた点

- 有用な能力は望ましくない能力にも一般化する（議論は割れている）
- 急激な能力向上は、2026年時点で未観測
- AIと人間の協働で科学が進む形は、Christianoの予測に近い

この投稿の結論：4年間の進展は、Paul Christianoの予測により好意的だった。

出典：LessWrong「Reevaluating “AGI Ruin: A List of Lethalities” in 2026」（個人の投稿）

別の見方｜doom派

もっと知りたい人へ

資料	読めること
Yudkowsky & Soares “If Anyone Builds It, Everyone Dies”（2025年9月）	論点の全体像。このスライドは概要記事と書評に基づく
Yudkowsky “AGI Ruin: A List of Lethalities”（2022年6月）	番号付きの項目で整理した、難しさの一覧
MIRI “MIRI’s Position on the Ban Artificial Superintelligence Act of 2026”（2026年9月）	禁止論の現在の政策的な立場
LessWrong “Reevaluating AGI Ruin in 2026”	4年後の検証。批判的な見方

まとめ

今週のまとめ

内部の脅威	スパイと策略するAIは、どちらも内部の脅威。違いは、AIのコピー同士が共謀しうること。
AIセキュリティ	重みの持ち出しを防ぐ（RANDのSL1〜5）。ジェイルブレイクは別のモデルにも転移する。
AIコントロール	策略するAIがいても、ローグ・デプロイメントや破局を防ぐ。ただし単純化した実験での結果で、能力差が広がると効きにくい。
doom派の論点	超知能は「育てられ」、一発勝負で止められないから、作らせない。ただし確率の連鎖への批判や、予測の検証もある。

任意資料

もっと知りたい人へ

AIコントロール

- **林央「AIを閉じ込めて見張れば安全？『AIコントロール』の難しさ」**（YouTube・約5分）懐疑的な立場からの解説
- **AGI Hub 用語解説「AIコントロール」**（日本語・短い定義）
- **CSET: AI Control: How to Make Use of Misbehaving AI Agents**（約16分）パート2が難しければ、まずこれ
- **Apollo Research: AI Behind Closed Doors**（内部展開のガバナンス）
- **Redwood ResearchのAIコントロール読書リスト**

AIセキュリティ

- **Lin et al.: AI Safety vs. AI Security**（2025年6月）セーフティとセキュリティを、意図の有無で区別する論文
- **XLab AI Security Guide**（シカゴ大学XLab）解説とコーディング演習で学べる教材