

人間の AI に対する判断移譲

現状と今後をめぐる調査報告

調査基準日 2026 年 10 月 21 日

GPT-6 Astra

要旨

AI への判断移譲は、作業手順の選択から、実行してよいかの判定、さらに人間へ確認を求めるかの判定へと広がっている。特にソフトウェア開発エージェントでは、毎回の人間承認に依存する方式から、権限の境界を先に定め、その範囲内を AI が進め、例外を人間に戻す方式への移行が確認できる。[1][3][4][5][6][7]

ただし、自律性が高いほど望ましいとは限らない。委ねられる範囲は、結果を検証できるか、失敗を取り消せるか、利害や価値観が対立するかによって変わる。人間と AI の組合せが単独の優れた判断者を上回る保証もなく、過信、技能低下、監督の形骸化について実証研究が蓄積している。

[11][12][15][16]

今後の焦点は、承認回数の削減だけではない。何を委ね、どの兆候で止め、誰が説明・修正できるようにするかを設計することである。技術上の判断移譲と、法律上・組織上の責任の移転は区別する必要がある。[18][19][20][21]

1 調査の範囲と資料の扱い

出発点は、西尾泰和氏の「人間の AI に対する判断移譲」と、指定された関連動画である。[1][2] 本報告は、この問題提起を、製品の公式発表、査読論文、研究機関の資料、制度・標準化の一次資料と照合して整理した。製品仕様は調査基準日時点の公表情報であり、プラン、設定、導入環境によって異なる。

指定動画の本文・字幕は取得できていない。そのため、動画全体を視聴・検証した要約としては扱わず、参考資料として書誌に収録した。[2] 将来見通しと実務設計は、本報告の分析・提案であり、既に確立した事実や制度とは区別して記す。

2 判断移譲には何が含まれるか

本報告では、判断移譲を「人間が担っていた選別、評価、選択、実行の可否、監督の配分を、一定の条件の下で AI に任せること」と捉える。単一の段階ではなく、次のように分解すると、何を委ねたのが明確になる。表 1 は本報告の整理である。

表 1 判断移譲の対象

対象	AI に委ねる判断	実務での例
情報選別	何を読むか、何を除外するか	関連特許やメールの選別

対象	AI に委ねる判断	実務での例
内容評価	正しいか、重要か、関連するか	技術的類似性や契約論点の評価
手段選択	どの順序・道具で進めるか	検索式、検索先、分析手順の変更
実行決定	操作を実行してよいか	ファイル編集、送信、申請、購入
監督配分	人間に確認を求めるか	例外の検出、承認依頼、停止
目的と価値	何を優先し、何を断念するか	費用と安全性の配分、権利の放棄

AI に最終的な実行権限がなくても、調べる資料、比較する選択肢、評価基準を AI が決めれば、人間の判断はその枠組みに影響される。人間が最後に承認したという事実だけでは、判断を実質的に保持したことにはならない。除外された情報や提示されなかった選択肢にも注意を向ける必要がある。

3 製品で進んだ変化

近年の変化は、単純な承認省略と、安全な範囲を定めた自動実行、別の AI による実行前審査を区別すると理解しやすい。承認画面が減ったという見かけが同じでも、制御の仕組みは異なる。

表 2 判断移譲に関係する主な公式発表

時期	公表された変化	意味と読み方
2024 年 12 月	Cursor 0.44 で Yolo Mode を導入。ターミナルコマンドの自動実行を可能にした。[3]	人間による都度確認を省く方向が製品化。
2025 年 10 月	Anthropic が Claude Code のサンドボックス化を説明。内部利用で許可プロンプトが 84%減少したと報告。[4]	実行環境の境界設定により、確認対象を絞る。
2026 年 3 月	Claude Code の auto mode を公開。操作を分類する仕組みで実行可否を判定する。[5]	確認すべき操作を選ぶ判断自体を AI が担う。
2026 年 4 月	OpenAI が Codex の Auto-review を解説。別のエージェントが権限拡張の要求などを審査する。[6]	作業を進める役割と承認を審査する役割を分ける。
2026 年 8 月	Claude Code で Pro、Max、Team の新規セッションを対象に auto mode を既定化すると発表。[7]	自動承認の仕組みが一部プランの標準的な開始状態へ。
2026 年 8～9 月	Claude in Chrome の一般提供と、OpenAI の dots を発表。[8][9]	ブラウザー等の日常業務にも、操作と確認の配分が広がる。

Claude Code の既定化は、2026 年 8 月 7 日の発表で 8 月 14 日からの新規セッションを対象としたものとされる。既存の明示的な設定や、対象外プランまで一律に変更されたという意味ではない。

[7] また、Codex の Auto-review は境界を越える要求等の審査を担う仕組みであり、すべての操作を独立した AI が再検証するものではない。[6]

4 人間の承認を増やすだけでは不十分な理由

承認の回数が増えると、一件ごとの注意が薄れ、内容を読まずに許可する運用になりやすい。

Anthropic は、有償のテスター1,053 人を用いた制御されたコーディング環境の評価で、危険なコマンドの検出率が人間では 13.6%、自動分類器では 89%だったと報告した。[7] これは同社の実験条件での結果であり、一般の職場やあらゆる判断で AI が人間を上回るという証拠ではない。

OpenAI も、自社の運用条件で Auto-review により人間の確認待ちによる中断が約 200 分の 1 になったと報告している。[6] この値が示すのは中断頻度の低下であり、安全性が 200 倍になったという意味ではない。AI が確認を省いた案件での見逃しや、事後の損害を別に測る必要がある。

利用者の行動も単純な全面委任ではない。Anthropic の観測では、経験の浅い Claude Code 利用者で完全自動に相当する利用は約 20%、経験を重ねた層では 40%超へと増える一方、経験者は実行中の介入も増やしていた。[10] この関連は、対象サービスの観測データに基づくものである。熟練した利用者は、自動化と能動的な監督を組み合わせている可能性があるが、因果関係までは確定していない。

したがって、監督を「毎回クリックすること」と同一視しない方がよい。目的と権限を設定し、途中の兆候を観察し、必要な時に止めるという役割まで含めて設計する必要がある。

5 実証研究から分かること

5.1 人間と AI の組合せは常に最良ではない

Vaccaro らのメタ分析は、106 件の実験と 370 の効果量を統合した。人間と AI の組合せは、単独の人間または AI のうち成績のよい方と比較すると、平均では劣っていた。効果量は Hedges の $g = -0.23$ 、95%信頼区間は $-0.39 \sim -0.07$ である。[11]

ただし、課題による違いがある。意思決定では損失が見られやすく、コンテンツ生成では相乗効果が得られやすかった。収集された研究の時期は主に 2020～2023 年であり、この結果を 2026 年のモデルの性能評価としてそのまま扱うことはできない。実務上の含意は、人間を追加すれば必ず品質が上がると仮定せず、役割分担を実測することである。[11]

5.2 批判的思考は確認と統合へ移る一方で弱まる可能性もある

Lee らは、知識労働者 319 人から生成 AI 利用の具体例 936 件を収集した。AI への信頼が高いほど批判的思考への労力が少なく、本人の能力への自信が高いほど労力が多いという関連が見られた。また、思考の中心が、情報の生成から、検証、回答の統合、作業全体の管理へ移ることが報告されている。[12]

これは自己申告による調査であり、AI 利用が知能低下を引き起こしたことを証明したものではない。しかし、AI 出力を評価する工程が残っていても、その工程へ実際に注意を向けるかどうかは、利用者の信頼や自信に左右されることを示唆する。[12]

5.3 誤った AI の助言を受け入れる認知的委譲

Shaw と Nave は、人が AI の回答に判断を預ける傾向を「cognitive surrender」として検討した。3 件の実験、計 1,300 人超、約 1 万試行を扱った研究について、Wharton の解説は、AI が正しい場合に正答率が基準条件より 25 ポイント上がる一方、AI が誤っている場合は 15 ポイント下がり、確信は高まり得ると紹介している。[13][14]

この研究は調査時点ではワーキングペーパーとして扱うべきであり、概念や効果の一般性を確立した定説ではない。それでも、正しい時の便益だけを測ると、誤った助言に引きずられる損失を見逃すという評価上の問題は明確である。[13][14]

5.4 人間と AI の相互作用で偏りが増幅する

Glickman と Sharot は、計 1,401 人を対象に、人間と AI の相互作用が知覚、感情、社会的判断をどう変えるかを調べた。偏りのある AI との相互作用を通じて、人間の判断の偏りが増幅し、本人がその影響を十分に認識しないことが示された。[15]

これは、AI の誤りが出力一件で完結せず、その後に人間が使う判断基準を変える可能性を意味する。人間による最終確認を置いても、確認者が同じ AI の提案に継続的に影響されていれば、独立した検証にならない場合がある。

5.5 作業中の成果と技能の獲得は別に評価する

Bastani らは、トルコの高校生約 1,000 人を対象に数学学習の無作為化比較試験を実施した。一般的な GPT 型支援を使った群は、練習中の成績が対照群より 48% 高かった一方、AI を外した試験では 17% 低かった。いずれも相対的な変化であり、点数の差を百分率ポイントで表した値ではない。[16]

答えの提供を抑え、学習を促す制約を設けた GPT Tutor では、この悪影響が緩和された。したがって、「AI を使うと学習が必ず悪化する」とも、「利用中の正答率が上がれば技能も育つ」とも言えない。設計と評価の対象を、即時の成果と独力での能力に分ける必要がある。[16]

6 期待できる便益と残る構造的なリスク

判断移譲が有効な場面では、処理速度、継続稼働、一貫した手順、確認できる範囲の拡大が便益になる。人間の注意を反復的な許可から重大な例外へ振り向けられる可能性もある。[4][6][7] ただし、実現した便益は業務ごとに測定する必要がある。

6.1 見えない除外と見えない不作為

実行された操作は記録しやすいが、AI が除外した資料、重要でないと判断した兆候、人間に知らせなかった案件は見落としやすい。監督を AI に委ねるほど、エスカレーションされた案件だけを見る運用では不足する。通知されなかった案件からも標本を抽出して点検することが重要になる。

6.2 正しさの確認と実行の許可を混同しない

計算や文書が正しいこと、ユーザーの意図に沿うこと、実行権限があることは別の条件である。例えば、正確な分析に基づくメールでも、送信先や機密区分を誤れば実行してはいけない。AI の評価器には内容の品質だけでなく、対象、権限、目的との整合を確認させる必要がある。

6.3 AI による二重確認にも共通の弱点がある

実行者と審査者を分けることは有用だが、同じ情報不足や誤った前提を共有すれば、両者が同じ判断を誤る。役割の分離は、誤りの独立性を保証しない。AI による審査に加え、権限の技術的制約、操作の記録、独立したテスト、異なる情報源を組み合わせることが必要である。[6][19]

6.4 自動化が進むと人間には難しい例外が残る

Bainbridge が 1983 年に示した自動化の問題は、日常の操作を機械へ任せるほど、人間が技能を維持しにくくなる一方、異常時には熟練した介入が要求されるというものである。[17] 生成 AI でも、通常案件を経験する機会が減ると、最も難しい例外だけを扱う人間の育成が難しくなる。

6.5 制御できない人に責任だけを集めない

Elish の「moral crumple zone」は、自動化された複雑なシステムで、人間の操作員が実際に持つ制御力に比べて過大な責任を負わされる問題を扱う。[18] 最終承認者に責任を置くなら、判断に必要な情報、検討時間、専門性、停止・修正の権限を伴わせる必要がある。

7 委譲する範囲をどう設計するか

Tomašev らの「Intelligent AI Delegation」は、単にタスクを分けるだけでなく、権限、役割、責任、監視、検証可能性を含む委譲の枠組みを提案している。[19] AI から別の AI への再委任も視野に入れた研究提案であり、実証済みの標準や認証制度ではない。

実務では、業務全体を一括して「自動化可能」と判断するより、判断の単位ごとに以下の条件を確認の方がよい。表 3 は本報告の提案である。

表 3 委譲範囲を決める評価軸

評価軸	委譲しやすい条件	人間の関与を厚くする条件
損害の大きさ	失敗の影響が小さく局所的	人命、権利、重要な資産に影響
取り消し可能性	復元ややり直しが容易	送信、公開、権利放棄など回復困難

評価軸	委譲しやすい条件	人間の関与を厚くする条件
検証可能性	正解や合否を外部に照らせる	根拠が不完全で評価が主観的
文脈の安定性	反復的で前提が明確	新規性が高く利害関係が複雑
価値観の対立	目標と優先順位が合意済み	利益配分、倫理、許容損失を選ぶ
技能の維持	独力で行う能力を別に育てられる	経験の蓄積自体が業務の目的

重大な業務でも、全工程を人間が手作業で行う必要はない。資料収集や形式検査は AI に委ね、影響の大きい判断点に人間の検討を集中させる設計が考えられる。逆に、低リスク業務でも、委譲範囲を AI が無断で広げることは防ぐべきである。

最低限、目的、利用可能な情報と道具、操作できる対象、費用・時間の上限、禁止事項、停止条件、人間に戻す条件、再委任の可否を明示する。実行の途中で前提が変わった場合には、最初の許可を無制限に引き継がない。

8 制度と標準化の現状

8.1 欧州の人間による監督

EU AI Act 第 14 条は、高リスク AI システムについて、リスク、自律性、利用状況に応じた人間の監督を求める。能力と限界の理解、自動化への過信への注意、出力の適切な解釈、無視・上書き・停止などが論点となる。[20]

これは、すべての AI 操作について人間のクリックを要求する規定ではない。何を理解でき、どの段階で介入できるかという実効性が重要である。本報告は、個別システムの法的分類や各条項の適用開始日の判定を行うものではない。

8.2 日本における民事責任の整理

経済産業省は 2026 年 4 月 9 日、「AI 利活用における民事責任の解釈適用に関する手引き」を公表した。人間の判断を AI が補助・支援する場合と、AI に依拠して判断を代替する場合などを区別して、既存法の解釈・適用を整理している。[21]

この手引きは、AI に判断を委ねれば利用者が当然に免責されるという制度ではない。組織では、現場の利用者、業務責任者、システム管理者、提供者の役割と、合理的に期待される確認を区別しておく必要がある。

8.3 エージェント間の接続と権限管理

NIST は 2026 年 2 月、AI Agent Standards Initiative を開始した。エージェントの相互運用性やセキュリティに加え、識別、認証、権限などに関係する標準化が課題になっている。[22] AI が他のサービスや AI へ作業を委ねるほど、誰の権限で何を行ったかを追跡できることが重要になる。

9 今後の見通し

以下は、現在の製品動向と研究から導いた本報告の見通しである。普及の速度や業務別の到達時期を確定的に予測するものではない。

表 4 進みそうな変化と確認すべき課題

変化	実務上の姿	残る課題
都度承認から条件付き委任へ	対象、金額、期間、通信先などを先に制限する	委任範囲が途中で広がることを防ぐ
出力の検査から監督判断の検査へ	何を人間に戻したか、戻さなかったかを監査する	通知されなかった重大案件の発見
単一 AI から役割を分けた AI へ	実行、審査、監視を別の役割として構成する	共通の誤りと責任分担
一律の自律性から案件別の調整へ	不確実性や影響度に応じて権限を変える	不確実性の自己評価の信頼性
操作記録から判断根拠の記録へ	参照資料、適用規則、権限の変更も残す	記録量、真正性、検証費用

AI の能力向上は委譲範囲を広げるが、ベンチマーク上の長時間タスク達成を、そのまま安全な無人運用時間と解釈してはいけない。METR の **time horizon** は、明確に定義された課題を人間の専門家が完了する時間に換算し、AI が一定の成功率で達成できる範囲を測る指標である。実業務の一日全体や、目標が曖昧な仕事の自律遂行を直接測っているわけではない。[23]

競争によって監督コストを下げる圧力が強まる可能性がある。その際は、処理時間と利用料金だけでなく、誤判断の損害、検証、復旧、教育、監査まで含めた総費用で評価する必要がある。承認が減ったことを成功の代理指標にすると、見逃しや技能低下を過小評価しやすい。

社会的な判断については、誰が目的と評価基準を決めるかが残る。Audrey Tang は、AI が人間に代わって社会を管理する方向への懸念と、市民の参加や多元性を保つ方向を論じている。[24] ここから導けるのは、判断を地域や組織の単位で限定し、異議申立てや再検討の経路を保持する重要性である。これは本報告による実務への解釈である。

10 知財業務に適用する場合

知財業務では、調査・比較・形式確認の委譲と、権利範囲・許容リスク・事業価値の判断を分けると導入を設計しやすい。表 5 は本報告の業務設計案であり、特定製品の能力保証ではない。

表 5 知財業務における役割分担の例

業務	AI に委ねる候補	人間が重点的に判断する事項
----	-----------	---------------

業務	AI に委ねる候補	人間が重点的に判断する事項
SDI と監視	分類、重複排除、変化の検出、重要度の一次評価	監視範囲、新たな兆候、見逃しの評価
先行技術調査	検索式の反復、候補の順位付け、対比の整理	調査目的、技術的範囲、調査を止める条件
FTO 調査	候補抽出、構成要件の対比、根拠箇所の整理	クレーム解釈、実製品との対応、リスク受容
出願書類作成	構成案、用語統一、記載の整合性確認	発明の本質、権利範囲、公開と秘匿の選択
拒絶理由対応	引用例の比較、反論案、補正案の列挙	権利と事業への影響、主張・補正の採否
ポートフォリオ	情報集約、費用整理、検討候補の抽出	権利放棄、出願国、投資、ライセンス方針

とりわけ注意すべきなのは、検索結果の中の誤検出だけでなく、検索・評価から外れた重要文献である。提示された候補の精度が高くても、重大な文献を取りこぼしていれば調査の目的を果たさない。採用候補と除外候補の両方を監査対象にする必要がある。

また、「これ以上調べなくてよい」という停止判断は、費用と見逃しリスクの配分である。AI が検索を繰り返せることと、目的に照らして調査を終えてよいことは別である。停止条件を事前に設定し、重要な前提が変わった場合は人間に戻す運用が望ましい。

11 導入と評価の進め方

導入時は、低リスクで結果を検証しやすい判断から始め、実測した失敗に応じて範囲を調整する。以下は、本報告が提案する最小限の運用手順である。

- 1 開始前に、目的、対象範囲、使用禁止の情報、権限、停止条件、人間へ戻す条件を決める。成功の定義には、処理件数だけでなく、重大な見逃しを含める。
- 2 実行中は、探索や手順変更を許す範囲を明確にする。前提の変化、根拠の不足、権限の拡張が必要になった時は、再評価を求める。
- 3 完了時には、結論だけでなく、根拠、除外理由、不明点、結論を覆し得る情報を確認する。AI による自己説明を、そのまま事実に基づく操作履歴とみなさない。
- 4 定期的に、採用された案件と除外された案件の両方を標本調査する。誤検出、見逃し、不要な確認、必要な確認の欠落を別々に数える。
- 5 人間の技能を維持するため、AI の提案を見る前に仮説や判断を作る機会と、AI を使わない評価機会を残す。教育段階では、答えを渡す支援と考え方を促す支援を使い分ける。[16][17]

評価指標には、時間短縮、重大な見逃し率、権限外の実行件数、人間に戻すべき案件の見逃し、復旧時間と損失、検証費用、AI なしでの技能を含める。単一の総合点へ早期に集約せず、どの便益と損失が増えたかを追える形を保つ。

人間が保持すべき役割は、すべての小さな操作を選ぶことではなく、目的と許容範囲を定め、その範囲が守られていることを確かめ、必要なら修正することである。その役割を実行できる情報と権限があって初めて、判断移譲は実効的な統制と両立する。

参考文献

本文の角括弧内の番号が、以下の文献番号に対応する。ウェブ資料の閲覧・照合基準日は 2026 年 10 月 21 日。DOI は恒久リンク、その他の URL は掲載ページへのリンクである。文献 2 は参照先としてのみ掲載し、内容を確認できていない。

- [1] 西尾泰和. 人間の AI に対する判断移譲. Scrapbox／Cosense, 2026 年 10 月 20 日. 指定された出発資料。
[https://scrapbox.io/nishio/人間の AI に対する判断移譲](https://scrapbox.io/nishio/人間のAIに対する判断移譲)
- [2] 投稿者・公開日未確認. 【激変】AI に作業だけでなく判断そのものを委ねる 2 年間の歴史. YouTube, ユーザー指定動画. 動画本文・字幕は未取得。内容を検証済みの根拠として使用していない。
<https://www.youtube.com/watch?v=dyHmz16ew-0>
- [3] Cursor. **Agent Improvements, Yolo Mode, Cursor Tab Update**. Changelog 0.44, 2024 年 12 月 17 日.
<https://cursor.com/changelog/0-44-x>
- [4] Anthropic. **Beyond permission prompts: making Claude Code more secure and autonomous**. Engineering／Research, 2025 年 10 月 20 日.
<https://www.anthropic.com/engineering/claude-code-sandboxing>
- [5] Anthropic. **How we built Claude Code auto mode: a safer way to skip permissions**. Research, 2026 年 3 月 25 日.
<https://www.anthropic.com/research/claude-code-auto-mode>
- [6] Trębacz, M., Arnesen, S., Matthews, O., Hurd, D., Park, W., Lin, O., & Gershenson, J.. **Auto-review of agent actions without synchronous human oversight**. OpenAI, 2026 年 4 月 30 日. 提供者自身によるシステム説明・運用評価。
<https://alignment.openai.com/auto-review/>
- [7] Phillippi, C.／Anthropic. **Auto mode is now the default in Claude Code for Pro, Max, and Team plans**. Claude, 2026 年 8 月 7 日. 対象プラン・新規セッション・既存設定に関する条件に注意。
<https://claude.com/resources/articles/auto-mode-default-in-claude-code>
- [8] Anthropic. **Claude in Chrome is generally available**. Claude Blog, 2026 年 8 月 26 日.
<https://claude.com/blog/claude-in-chrome-generally-available>
- [9] OpenAI. **Introducing dots**. 2026 年 9 月 29 日.
<https://openai.com/index/introducing-dots/>
- [10] Anthropic. **Measuring AI agent autonomy in practice**. Research, 2026 年 2 月 18 日. サービス利用の観測データに基づく分析。
<https://www.anthropic.com/research/measuring-agent-autonomy>

- [11] Vaccaro, M., Almaatouq, A., & Malone, T. W.. **When combinations of humans and AI are useful: A systematic review and meta-analysis.** Nature Human Behaviour, 8, 2293–2303, 2024. DOI: 10.1038/s41562-024-02024-1. 査読論文。106 実験、370 効果量のメタ分析。
<https://doi.org/10.1038/s41562-024-02024-1>
- [12] Lee, H.-P. (H.), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N.. **The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.** Proceedings of CHI 2025, Article 1121, 1–22. DOI: 10.1145/3706598.3713778. 査読会議論文。319 人、936 件の具体例に基づく自己申告調査。
<https://doi.org/10.1145/3706598.3713778>
- [13] Shaw, S. D., & Nave, G.. **Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender.** Working paper, 2026 年 1 月 11 日. DOI: 10.2139/ssrn.6097646. ワーキングペーパー。本文での数値説明は文献 14 も参照。
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6097646
- [14] Wharton Executive Education. **Thinking Fast, Slow, Artificially: AI and Your Brain.** Wharton at Work, 2026 年 5 月. 文献 13 を紹介する研究機関の解説。
<https://executiveeducation.wharton.upenn.edu/thought-leadership/wharton-at-work/2026/05/thinking-fast-slow-and-artificially/>
- [15] Glickman, M., & Sharot, T.. **How human-AI feedback loops alter human perceptual, emotional and social judgements.** Nature Human Behaviour, 9, 345–359, 2025. オンライン公開 2024 年 12 月 18 日. DOI: 10.1038/s41562-024-02077-2. 査読論文。計 1,401 人の研究。
<https://doi.org/10.1038/s41562-024-02077-2>
- [16] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., & Mariman, R.. **Generative AI without guardrails can harm learning: Evidence from high school mathematics.** Proceedings of the National Academy of Sciences, 122(26), e2422633122, 2025. DOI: 10.1073/pnas.2422633122. 査読論文。高校数学の無作為化比較試験。
<https://doi.org/10.1073/pnas.2422633122>
- [17] Bainbridge, L.. **Ironies of automation.** Automatica, 19(6), 775–779, 1983. DOI: 10.1016/0005-1098(83)90046-8. 自動化と人間の技能・監督に関する基礎論文。
[https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- [18] Elish, M. C.. **Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.** Engaging Science, Technology, and Society, 5, 40–60, 2019. DOI: 10.17351/ests2019.260.
<https://doi.org/10.17351/ests2019.260>
- [19] Tomašev, N., Franklin, M., & Osindero, S.. **Intelligent AI Delegation.** arXiv:2602.11865, 2026 年 2 月 12 日. プレプリント。委譲の設計枠組みを提案する研究。
<https://arxiv.org/abs/2602.11865>
- [20] European Commission. **Article 14: Human oversight.** AI Act Service Desk, 2026 年 7 月 27 日統合版に対応する条文掲載。高リスク AI に関する人間の監督の規定。
<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-14>
- [21] 経済産業省. **AI 利活用における民事責任の解釈適用に関する手引き.** 第 1.0 版, 2026 年 4 月 9 日公表. 既存の民事法の解釈・適用を整理した公表資料。
<https://www.meti.go.jp/press/2026/04/20260409001/20260409001.html>

- [22] National Institute of Standards and Technology. **AI Agent Standards Initiative**. NIST, 2026 年 2 月 17 日公表, 2026 年 8 月 14 日更新.
<https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>
- [23] METR. **Task-Completion Time Horizons of Frontier AI Models**. 2026 年 9 月 8 日更新の公開資料. 人間の専門家の所要時間を基準としたタスク達成指標.
<https://metr.org/time-horizons/>
- [24] Tang, A.. **Plurality & 6pack.care**. 著者による講演記録, LessWrong, 2025 年 9 月 1 日掲載. DeepMind での講演に関する本人公開資料。日付は掲載サイト表記。
<https://www.lesswrong.com/posts/anoK4akwe8PKjtzkL/plurality-and-6pack-care>