Deloitte.
デロイト トーマツ

AIのモラルライセンスについて考える

地域社会と対話する上での道徳や倫理の必要性

Deloitte Australia Centre for the Edgeについて

Deloitte Australia Centre for the Edgeは、デロイトのグローバル研究ネットワークの一部であり、新たな技術の機会を通じて企業が利益を生むための支援を行います。技術開発は企業の根幹をなすものであり、企業の成長戦略を見出すことが重要です。私たちのミッションは、シニア・マネジメントのアジェンダにはまだなっていないがアジェンダとなるべき機会を特定し、調査することであり、長期的なトレンドと機会に焦点を当てながら、短期的な行動への影響にも注目しています。

企業の皆様には、当センターの国際的な研究に加え、現地のビジネス環境に影響を与える問題に関する研究に関する知見を提供しています。

連邦科学産業研究機構（CSIRO）について | Data 61

オーストラリア連邦科学産業研究機構 (CSIRO) は、オーストラリアの国立科学研究機関です。CSIROでは、革新的な科学技術を用いて課題の解決に取り組んでいます。CSIROは未来を形作ります。CSIROは科学の力を利用して現実の問題を解決し、私たちの地域社会、経済、地球にとってより良い未来を開きます。

私たちの世界は急速に変化しており、データはこの新しい世界の基本通貨です。CSIROのData 61はオーストラリアの主要なデジタル研究ネットワークです。お客様のデータドリブンな未来の構築を支援します。

Deloitte AI Institute (DAII) について

Deloitte AI Institute（DAII）は、AIの戦略的活用およびガバナンスに関する研究活動を行うと共に国内外のAI専門家とのネットワーク形成を行う研究組織です。デロイト トーマツの様々なビジネスの専門家と連携することで、研究成果をベースに日本企業のAIによるビジネスの変革と持続的なオペレーションを支援していきます。日本のDAII設立により、海外の約6,000人のAI専門家の知見の活用や、各地のAI専門家コミュニティとの交流を行っています。これにより、研究活動のスピードや質の向上、海外の多様なユースケースや実証実験の経験の共有を促進していきます。

目次

AIと倫理：社会的営業免許（Social License）の問題とは？

「私の会社は、地域貢献活動に年間700万ドルを費やしています。しかし、未だに私たちは支援する地域から反対を受け、仕事を中断せざるを得ません。明らかにお金では必要な善意は伝わっていないことがわかっていますが、我々は地域の声のどこを見落としているのかわからないのです。」[1]

—石油会社のマネージングディレクター

3人の友人は、オーストラリアのニューサウスウェールズ州（ニューサウスウェールズ州）のノーザンリバーの農場で朝のお茶を飲んでいたとき、谷の反対側にある隣人の敷地に掘削装置が設置されているのに気がつきました。それまで、彼らはコールシームガス（Coal seam gas: CSG）産業のことを聞いたこともなければ、それまでCSGに対する反対活動を考えたこともありませんでした。しかし、その掘削装置を現実に目の前にした時、彼らを反対運動へと駆り立てるには十分だったのです。このグループはすぐに反CSG運動の確立に貢献するようになりました。その活動は広がり、ついには2014年にニューサウスウェールズ州政府がこの地域でのガス探査許可を停止する結果となったのです[2]。さらには2015年までに、政府はこの地域の50万ヘクタールに及ぶ石油探査ライセンスを買い戻す事態になりました[3]。

鉱業会社は、多くの産業界の企業と同様に、事業を行うための「合法的なライセンス(営業許可)」を持つことと、「道徳的なライセンス」を持つことの違いに悩まされてきました[4]。言い換えると、「できること」と「すべきこと」の区別であり、技術的に可能で経済的に実現可能なことだからといって、その影響を与える人々が道徳的に受け入れられるとは限らないのです。つまり、地域社会に受け入れられなければ、企業は「会社は私たちのために何もしてくれなかった」という「地域のトラブルメーカー」からの「終わりのない要求」に直面することになり、結果として、財政的、非財政的なコストがプロジェクトを圧迫することになるのです[5]。結果的に、企業が善意で（と思われるものに）投資していても、冒頭にあった事例のように地域社会からの反対を受けることもあるのです。企業は地域の風潮を理解し、地域の社会インフラ（医療や教育へのアクセスを改善し、道路や電力サービスを改善し、地域の経済活動を促進して地元企業の活気と健全な雇用市場を生み出す）に投資することもあるかもしれませんが、結局のところ効果がない場合が多いのです。

コミュニティからの受け入れがなく、「道徳的なライセンス」がないため、ニューサウスウェールズ州の鉱山会社は苦労していました。この道徳的なライセンスは一般的に社会的営業免許（Social License）と呼ばれる90年代に作られた造語で、地域社会が鉱山開発を継続的に受け入れ、承認することを表しています。それ以来、企業は地域社会と協力して社会的操業許可（social license to operate、SLO）を取得し、維持しなければならないことが鉱業内でますます認識されるようになりました[6]。操業する社会的営業免許の概念は時間とともに発展し、伐採や紙パルプ工場など、操業する物理的環境に影響を与えるさまざまな業界で採用されています。

先の事例は、人工知能（AI）とどのような関係があるのでしょうか？AIは、採掘、伐採、紙の生産からは遠く離れているように見えるかもしれませんが、AIを扱う組織や企業（最近ではほとんどの企業がそうだと思われます）は、テクノロジーの利活用が社会に受け入れられ、社会に影響を与え、その受容と影響に関しているという点で、同様の課題を抱えていることに気付いています。AIソリューションがどれだ

け慎重に設計されていても、どれだけユーザーグループによるテストが行われていても、一般の人にソリューションを公開すると、さまざまなバイアスを含めた反応がどうしても発生してしまします。Bluetooth対応の健康器具は賛否両論がありますが、ある人はこのソリューションを、恥ずかしさや健康上の問題を回避するのに役立つ恩恵と見る一方で、プライバシーや安全性の懸念や、デバイスがハッキングされて個人情報が漏れることを心配する人もいます[7]。例えば、刑事裁判で被告人の再犯リスクを推定するツール[8]であるCOMPAS, [9] や、詐欺を検出して自動的に不実表示で告発し、返済を要求することを目的としたソリューションであるMiDAS, [10] については現在もなお議論されております。なぜならば、これらのソリューションは、恵まれないグループに偏っており、社会の構造的不平等を悪化させ、不利益を制度化していると考えられているのです。石油掘削装置の建設と同様に、AIソリューションが法的にも経済的にも実現可能であるという事実は、たとえ個人的に利益を得る立場にあったとしても、地域社会がそれを道徳的または倫理的に受け入れられることを意味するものではありません。

AIは万能ではありません。すべてのテクノロジーと同様に、個人や社会全体に利益をもたらすこともあれば、害を与えることもあります。テクノロジーをどのように使用するか、つまりテクノロジーをアイデアからソリューションに変換する方法によって、潜在的なメリットが害を上回るかどうかが決まるのです。「テクノロジーは良くも悪くもありません。」[11] どちらも熟考が必要なため、テクノロジーをどのように使用するか、何を目的として、どのような手段でテクノロジーを使用するかが重要です。問題を最小限に抑えるか、適切に管理しながら、利益を実現するためには、選択と妥協が必要です。潜在的な問題のために技術を放棄することは、最も望ましい選択肢ではないかもしれません。というのは、（すでに）不完全な世界における「十分に良い」解決策は、バランス上、不完全な世界そのものよりも好ましいかもしれないからです。しかし、問題は「十分に良い」とは何かということです。

では、見極めるためには何をすべきなのでしょうか。どのようにしてこれらの機会を特定すればいいのでしょうか。そして、妥協するためにはどのようなプロセスを使えばよいのでしょうか。地域社会内の多様な声に耳を傾け、彼らの懸念を説明できるようにするにはどうすればよいのでしょうか？

課題を構造化する

AIシステムは、機械翻訳、自動運転車、音声アシスタント、文字と手書きの認識、広告ターゲティング、製品のレコメンデーション、音楽認識、顔認識など、さまざまなソリューションを実現することができます。その中でAIは指示、アドバイス、測定の報告、情報と分析の提供、実行された作業の報告、自身の状態の報告、シミュレーションの実行、仮想環境のレンダリングに使用されています[12]。数年前には不可能と思われていたソリューションが、今では私たちが毎日使う製品やサービスに組み込まれているのです。

AIに対する私たちの見方も変わりました。AIを動力とする解決策が人間の弱点の一部に対抗するという期待は、AIが実存的な脅威であるかもしれないという恐れに変わったのです。最初は、規制によってAIがどのように使用されるかを制御できると考えられていました[13]。公開書簡が政府に送られ、署名者の長いリストが添付され、規制の制定を求められましたが[14]、これは実を結びませんでした。最近では、AI対応ソリューションの開発を導くための倫理原則の開発に焦点が当てられています。これらの原則は、私たちがAIに何を求めているのか（何を避けたいのか）を抽出した有用なものですが[15]、特定のソリューションがどのようにそれらを遵守すべきかを説明するには至らず、十分ではありません[16]。最近では、デザイン（およびデザイン方法論）によってこれらの原則を適用できるようになることが期待されていますが、デザインだけで十分であるかどうかは定かではありません。

問題を最小限に抑えながらAIの価値を実現するという課題に取り組むためには、次の3つの論点によって複雑になっています。

- AIとは何であるか、したがって問題が何であるかを理解するという定義上の課題
- 技術（AI）ソリューションを社会規範に合わせるという課題
- 異なるコミュニティ同士（社会的世界）の橋渡しをするという課題[17]—構成員が現実世界をどのように理解し、考えているかを形作る、社会における異なる文化的セグメントを繋ぐ

本稿では、上記の論点をそれぞれ順番に議論していきます。

定義上の課題：AIとは何か？そして問題は何か？

AIとは何か、また何ではないかについて、広く合意された正確な定義はありません。これは、AIが幅広い技術の集合体であり、無関係な様々なテクノロジーも含まれるためです。しかし、実用的な定義は次の通りとされています。

「人工知能とは、機械を知能化するための活動であり、知能とは、環境の中で先見性を持って適切に機能することを可能にする資質のことである。」[18]

—Nils J. Nilsson

不正確ですが、この定義は、私たちがAIプロジェクトと呼ぶ可能性のある膨大な範囲を捉えています。正確な定義がないことも、この分野の成長を助けているかもしれません。なぜならばAIは、その実践者が目標を追求するために他の分野からアイデアや技術を「借用」することで[19]、一種の装飾建築物を建てることを可能にしているからです[20]。より皮肉な言い方をすれば、AIは「まだ完全には機能していないもの」と定義できるかもしれません[21]。そして一方でAIが広く採用されると、多くのテクノロジーがAIと見なされなくなります。ロボット工学者のRodney Brooks氏[22]はかつて、次のように不満を述べていました。「その一部を理解するたびに、それは魔法ではなくなる：『ああ、単なる計算だ』と言うのだ」[23]AIは（現時点では）不可能なことを示すラベルである、という感覚があるのです。

AIテクノロジーは、AIコミュニティが興味深く感じる問題を解決するために使用するテクノロジーです。つまり、単に物理的ではない、人間の認知的成果を再現するために活動しているコミュニティと見なすことができます。実際、AIに見られる現在の投資の重要な推進力は、クラウドサービス、データへの簡単なアクセス、低コストのユビキタスコンピューティングとネットワークであり、それ自体が新しい破壊的な技術の開発ではなく、古い技術から新しいソリューションを構築することを主眼に置かれています[24]。数十年にわたる着実な進歩の後、新しいAI技術の発見は一見、停滞しているように思われています[25]。

「知的な」技術とそれ以外の技術の境界線をどこで引くかに関わらず、AIの倫理に対する懸念が高まっているのは、CRISPR [26]や遺伝子組み換え生物（GMO）の開発のように、これまでにない新しい技術のせいではありません。AIが懸念される倫理的な課題は、テクノロジーの独自の機能によるものではなく、テクノロジーを大規模に簡単かつ安価に展開できることによるものです。破壊的なのは、この展開の規模です。技術史家のMelvin Kranzberg氏は次のように述べています。

技術に関連した問題の多くは、一見良さそうに見える技術が大規模に使用された場合に、予期せぬ結果を招くために発生します。したがって、最初に導入されたときには人類に恩恵をもたらすと思われた多くの技術的応用が、その使用が広まったときには脅威となったのです[27]。

AI展開の規模が拡大しているおかげで、社会は転換点にあるようです。いくつかの自動化された意思決定による世界から、多くが自動化された意思決定による世界への移行です[28]。社会は意思決定をアルゴリズムで形式化、それをソフトウェアで固めて自動化し、それらの意思決定を相互に、そしてそれらを取り巻く運用ソリューションと結びつけているのです[29]。以前のデジタルは、エンタープライズ・アプリケーションとパーソナル・コンピューティングで構成されていましたが、現在では、常にオンラインで利用可能で、クラウドソリューションとスマートフォンで相互に接続されるようになりました。

自動化された意思決定は現実世界に影響を与えるハードウェアと統合しています。そして、重複する意思決定ネットワークによって支配されている状況を作り出しています[30]。そのため、自動化されている個々の意思決定それ自体が必ずしも問題になるわけではありません。むしろ、問題のある行動は、個々の意図しない結果が、自動化された意思決定の間で衝突が起こり、「スマートな」システムがうまくいかなくなるような状況が発生することで、自動化された意思決定がうまく統合されず、顕在化するのです[31]。

例えば、レンタカー会社は、Internet of Things [32](IoT)センサーやエフェクター[33]を利用して、支払いから提供までのエンドツーエンドのレンタルプロセスを、個々のレンタカーに至るまで統合することができるかもしれません。これにより、企業は車の位置を追跡し、よりカスタマイズされたレンタルプランを提供し、路上での レンタカー利用者をサポートすることが可能になる可能性がある一方で、万が一車が盗まれた場合には車の機能を停止することで盗難を減らすことができます。しかし、これらのシステムは、レンタカー利用者が携帯電話の電波が断続的に届く遠隔地でキャンプをしているときに、会社がSMSやアウトバウンド・コールセンターを介してレンタカー利用者と連絡が取れなくなったときに、支払いゲートウェイに一時的な障害が発生したために車が盗まれたと判断され、レンタカーを停止してしまう可能性があるのです。借主は、外に出て助けに連絡したりすることができず、車両の中で放置されることになります[34]。

重要なのは、悪い（自動化された）決定が一連のノックオン効果を連鎖的に引き起こし、問題をエスカレートさせ、さらに悪い決定を引き起こす可能性があるということです[35]。Kranbergが警告する予期せぬ結果とは、以前に手動で行った意思決定が自動化されて統合された後、意図しない相互作用がなされることなのです。加えて、これらの相互作用は、レンタカーの例のように、非常に偶発的なものである可能性があります。また、会社の合併後に誤って解雇リストに名前を追加してしまい、従業員を解雇して再雇用せざるを得なくなるなどの事例もあり得ます [36]。給与計算を運用管理システムやアクセス管理システムと統合することで、内部プロセスが合理化されるだけでなく、自動化された意思決定のネットワークが形成され、一度開始されると、もはや会社がコントロールすることはできません。

そのため、私たちは次の4つの「領域」を考慮する必要があります：[37] ① 私たちは正しいことをしているか、② 私たちはそれらを正しい方法でやっているか、③ 私たちはそれらをうまく作動させているか、④ そして私たちは有益を得ているか。そしてその上で、これらの自動化された意思決定を展開および統合することで、ガバナンスと監視が低下し、プライバシーや利用者の同意に関する問題について考える必要が前面に出てくるのです。

そのため、AIの道徳的ライセンスを考える上では、個別のテクノロジーではなくシステムに焦点を当てる必要があるのです。

技術的ソリューションを社会的規範に合わせる

私たちの2つ目の課題は、技術的（AI）ソリューションを社会的規範に合わせる問題です。技術的なコミュニティは、その分析的アプローチの性質上、細部に焦点を当てています。自律走行型の車を作る問題は、赤信号に近づいたときにどうするか、歩行者が車の前でつまずいたときにどうするかなど、様々な状況下で車がどのように振る舞うべきかを定義します。「正しい」車の動作を設計することは、十分に異なるコンテキスト（異なる動作シナリオ）を特定し、各状況に適切な応答を作成することです。同様に、偏りのない顔認識アルゴリズムを作成するには、アルゴリズムを設計するために使用される行動シナリオ（および応答）のセットが、過去の（そして偏った可能性のある）データセットに頼るのではなく、人口統計的にバランスのとれた画像セットで訓練された、適切に偏りのないものであることを確認する必要があると考えられます。

特定の反応が倫理的であるかどうか（またはそうでないか）は、しばしば「文脈に依る」問題であるため、この還元主義的アプローチは当然問題と見なされます。自動運転車の場合、これはトロッコ問題に現れます。これは、1967年にPhillipa Foot 氏[38]によって現代の形で最初に提起された思考実験です[39]。トロッコ問題は、人間のオペレーターが、トラックを変更するレバーを引くかどうかを選択する必要があるというジレンマを扱ったものです。複数人のグループがあるトラックに立っているのに対し、別の個人がもう一方のトラックに立っているため、オペレーターは、自身が行動を起こさずに死ぬか、自身が行動を起こして別の個人が死ぬかを選択するかを問う問題です。ここでのポイントは、単一の「正しい」選択はないということです。選択は、特定の状況に適用される主観的な値に基づいて行われ、選択を拒否することもできません[40]。自動運転車で特定されたシナリオの多くは、明確な解答がなく、合理的な個人が同意しない可能性があります。適切な対応は、特定のシナリオに対するものです。同様に、顔認識システムのトレーニングセットを人口統計に合わせようとすると、どのグループの人々が使用する人口統計プロファイルを決定するかという問題が発生します。

多様で複雑な現実世界では、倫理的行動を確実にするために、どのような問題でも十分な数のシナリオに切り分けることが、果てしなく無駄な作業になります[41]。新たに定義されたシナリオは、既存のシナリオと相反する可能性がありますが、これは主に、これらのシステムが、その性質上、流動的で不正確な人間が定義した（社会的に決定された）カテゴリーやタイプを使って作業を行っているからです。また、ソリューションの動作コンテキストを変更すると、人口動態や環境の性質に関する仮定が適用されるシナリオが適用されなくなるため、シナリオを検討するために費やしてきた努力がすべて台無しになる可能性があります。例えば、ヨーロッパで設計された自律走行型自動車は、オーストラリアの野生生物によって混乱する可能性があります[42]。あるいは、医療診断のソリューションは、実験室では成功しても現実世界では失敗するかもしれません[43]。

自然なバイアスは、「公正」または「倫理的」な状態をアルゴリズム的に定義できると考えられますが、これは不可能です[44]。なぜならばこれは、一般的には、技術者にとっての盲点となり、全て対応することは不可能であるからです。

異なるコミュニティの架け橋となる

3つ目の最後の課題は、異なるコミュニティの架け橋となることです。私たち全員が独自の生きた経験、私たちが誰であるか、そして私たちが世界や社会にどのようにアプローチするかを形作った個々の歴史を持っています。1930年代の大恐慌の時代に生まれた世代がその一例です。その間に破綻した銀行は、数え切れないほどの個人の人生の貯蓄を彼らと一緒に奪い、多くの人々の間で銀行に対する生涯の不信感を生み出しました。

社会における意見の相違は、一般的に価値観や原則の違い、つまり私たちの周りにあるものをどのように評価するかの違いとして捉えられています。社会の不一致は、通常、価値観や原則の違い、私たちの周りにあるものを評価する方法の違いとして組み立てられます。しかし、社会の最も深くて手に負えない論争のいくつかは、主に価値観と原則に関するものではありません。確かに、私たちはしばしば原則に同意することができます。違いは、私たちがこれらの価値観と原則を適用するコミュニティ、つまり私たちの周りにあるものを解釈する方法にあります[45]。例えば、「人を（不当に）殺すことは間違っている」という原則には同意しても、何が人を構成するのかについては意見が合わないことがあるかもしません[46]。

このような最も難解な論争の進展は困難です。なぜなら、公平性のような原則を測るための基準として、完全に統一化されたコミュニティ[47] が存在すると仮定するのが一般的だからです。この仮定は、誰もが自分たちと同じ世界を見ているが、異なる価値観でそれに近づこうとしているだけで、それが必ずしもそうではないというものである[48] というもので、多くの社会評論家の盲点となっているのです。

このようなコミュニティ間の違いが、最近、物議を醸しているAIソリューションの中にも表れているのを見ることができます。前述の再犯予測ツールであるCOMPASは良い例です。COMPAS を開発したチームは、すべての個人が平等に扱われ、最も多くの人にとっての害（大まかには、誤った予測の割合）が最小になるような世界のためのソリューションを作成するという、実用主義[49] のアプローチをとりました。異なる尺度を使用し、異なる世界の規範に従ってCOMPASを判断する場合、その世界は、すべての個人がどのような状況からスタートしても、人生において同様の結果を経験するという公平性に焦点を当てたものです[50] 。とすれば、COMPASが引き起こす意図しない危害は不利益を被ったグループに不釣り合いとなるため、COMPASは欠落しているものがあると言えます[51]。これは「公平性のパラドックス」[52] であり、あるコミュニティでCOMPASのパフォーマンスを向上させると、他のコミュニティではパフォーマンスが低下します（逆もまた同様です）。

AIソリューションは倫理的である必要があり、公平性（公正な扱いと結果の促進）や危害の回避などの原則に従う必要があることに同意します[53]が、これらの原則を実践に移すためにどのトレードオフが必要か、原則がどのように定められるかについても意見が分かれます。明確に定義された同じ原則をさまざまな社会の世界に適用すると、結果が大きく異なる可能性があります。そのため、オープンで多様な社会では、同じ原則のセットで作業する様々なチームが全く異なるソリューションを作成する可能性があります。これらの違いにより、あるグループが別のグループのソリューションを非倫理的であると考える可能性があります。

カンファレンスで（修辞的な）質問をするのはよくあることです：何が倫理的であるかは誰が決めるのでしょうか？設計上の決定は、一部の人口統計グループの権利を剥奪したり、特定のグループに影響を与える可能性があり、既存の不平等や不利益に対処できない可能性があります。そのため、適切に敏感な意思決定者が決定を下すように注意する必要があります。しかし、誰が意思決定を行うかに焦点を当てるということは、この個人の特定のコミュニティ（何が倫理的であるか、倫理的でないかの枠組みに使用される）がどのように選択されたかを無視していることを意味するので、これは間違った質問になりそうです[54] 。特定のソリューションが触れる異なるコミュニティ間に、どのようにして橋を架けることができるのでしょうか？トレードオフはありますが、橋渡しがなければ、どのようなトレードオフを設けるかを決めることはできません。

倫理的な AI ソリューション（道徳的意思決定ネットワーク）を開発する際のジレンマを要約すると、勝てない、損切りできない、ゲームから離れることができないということになるかもしれません[55] 。単一のコミュニティ、つまり想定される世俗的な社会の観点から「倫理的」を組み立てることを選択した場合、そのコミュニティを他のコミュニティよりも優先しなければならないため勝つことはできません。たとえ中間地点、コミュニティ間の架け橋を見つけることができたとしても、私たちの技術的ソリューションは、私たちが非倫理的だと考える可能性のある問題や例外に溢れているからです。また、私たちが経験しているのは、独立した自動決定を含む世界から、そこに含まれる相互作用する自動決定のネットワークによって定義される世界への移行であるため、望ましくないテクノロジーを禁止または規制してゲームを離れることはできません[56]。

現在の膠着状態を乗り越えるには、これらすべての課題に対処する方法を見つける必要があります。それは、（他の人よりも一方を優先するのではなく）関与するすべてのコミュニティの懸念に対処することを可能にする方法であり、（提案された）システムと（技術だけではなく）それが触れるコミュニティの両方を考慮する方法であり、また、どのような自動化された意思決定システムにも内在する葛藤や不確実性、倫理的な欠陥を管理するためのメカニズムを提供する方法でもあります。私たちは、包括的な対話を必要としているのです。

信頼と受容

成功したAIソリューション、つまり成功した自動意思決定ネットワークとは、意図された機能を効果的に実行するだけでなく、それに触れる人々に受け入れられ、承認され、最終的に信頼されるものです[57]。課題がある場合、経営者は「地元のトラブルメーカー」からの「終わりのない要求」に対応したり、「会社は何もしてくれない」と言われたり、プロジェクトを圧迫するコストを負担していることに気づきます。経営者とコミュニティ[58] の関係は、敵対的なものではなく、協力的なものでなければならず、AI をどのような時に使うべきかを理解するために協力し合う必要があります。残念ながら、私たちはそのような状態には程遠いのが現状です。

AI のための社会的操業許可（Social License to Operate、SLO）の概念は、上記で説明した3つの課題（定義、ソリューションを社会規範に合わせる、社会的世界を橋渡しする）すべてに対応する可能性を秘めています。SLOは、テクノロジーではなく、ソリューション全体とそれが展開される社会的および物理的環境に焦点を当て、特定のAIテクノロジーにメソッドを集中させることで問題を回避できます[59]。SLOはまた、コミュニティ間の橋渡しとして、ソリューション自体が共通の倫理的な見解と考えられないことを認めることが課題の対処になります[60] 。企業は法的に事業を行う権利を持っているかもしれませんが、その上でコミュニティの同意を得て事業を行うための道徳的なライセンスを取得しなければならず、このライセンスは、ソリューションとコミュニティの両方が進化し、状況が変化するにつれて維持され、更新されなければなりません[61]。SLOを開発および維持する継続的なプロセスにより、企業は、影響を受けるコミュニティ間に架け橋を築くことができます。このSLOプロセスによって、会社がコミュニティと協力して、互いに提案されたソリューション、そして各当事者の目標、規範、原則を理解するための枠組みを提供することができます。そして、提案されたソリューションについての共通理解を深め（特定の技術ではなく意思決定ネットワークに焦点を当てて）、共有された原則が実際の生活の中でどのように実行されているかを判断し、問題と機会を特定して解決策を見つけることで、ソリューションを社会規範に合わせるという問題に対処できるのです。対話がオープンで包括的なものであるためには、正直さが求められるため、脆弱であることを意味します。もちろん、製品やサービスは公平に表現されなければならないし、利害関係者は、技術が誤った表現をされていないこと、あるいはインフォームドコンセントがあった場合には、データが約束通りに保存され、使用されることを信頼する必要があります。

対話がオープンで包括的なものであるためには、正直さが求められるため、脆弱であることを意味します。

ケーススタディ：インテリジェントな病院

ある企業が「スマート」な病院を開発している場合を考えてみましょう。この病院は、スマートビルの設備をすべて備えています。住民が建物をどのように使用しているかを追跡するIoTセンサーネットワーク（部屋の使用パターンと個人の好みを特定）と、建物の運用を最適化して個人に合わせて調整し、メンテナンスコストを最小限に抑える自動化設備などです。建物の環境フットプリントを削減し、ユーザーの利便性と快適性を向上させます。フロアごと、ゾーンごとの空気の質とスタッフの有無をデータ化することで、空調と暖房を最適化し、電力と水の使用量を減らしながら快適性を向上させることができます。周囲の光のレベルとスタッフの活動に関するデータを使用して、照明を最小限に抑えることができます。バックアップ発電機や酸素供給ラインなどのプラント機器を計測することで、ジャストインタイムのメンテナンスが可能になります。スマートフォンのアプリを利用することで、利用者はこれらのシステムと対話し、自分の体験をパーソナライズすることができるようになります。

AIを使用してこれらのシステムをつなぎ合わせると、スマートな病院が「インテリジェント」な病院に変貌します。音声アシスタントはどこにでも設置され、登録（救急室を含む）、患者・治療室、手術室などに設置され、スタッフ、患者、来客が病院のプロセスをより便利に操作したり、助けを求めたり、言葉の壁を乗り越えたりすることができるようになります。スタッフ、患者、来訪者は、建物に初めて入った時から追跡され、業務システムに保存された記録と関連づけられています。患者は二度と行方不明になることはありませんし、来訪者は道案内され、スタッフは緊急時にはいつでも最寄りの専門医を見つけることができます。意思決定支援ツールは診断を迅速化し、医療画像上で潜在的な問題を強調表示し、患者の特定の症状の集合が何を暗示しているかを示します。これらの情報はすべて、AIを搭載した状況認識および計画システムに入力され、（おそらく緊急事態に発展する前に）問題を特定し、意思決定者に潜在的な問題と解決策の両方を提示します[62]。その結果、状況認識・計画システムはドローンのクラッシュカートを派遣し、サポートスタッフと最寄りの専門医に警告を発し、潜在的な緊急事態に対応するために手術室のスケジュールを変更することを提案します。

恩恵はあるものの、このインテリジェントな病院は、大規模なAIの展開に関連する問題の多くに対応する可能性があります。例えば、音声アシスタントは様々な言語をサポートしなければならなりませんし、そのソリューションに偏りが出ないように各言語内のどの方言をサポートすべきか[63]、そして（何らかの理由で）話せない人を病院はどのようにサポートすべきか検討が必要です。X線画像を「読み取り」、肺損傷やその他の肺炎の兆候を強調するツールは、先進的な病院ではうまく機能していましたが、インテリジェントな病院がサービスを提供している人口統計学的グループの1つに偏っている可能性があり、偽陰性または陽性の結果が出る可能性は否定できません。特定の専門家や機械などの希少なリソースに対する相反するニーズに直面した場合、状況認識とシステムの計画とどちらを優先させるべきでしょうか。どの患者が優先されるのか、また、システムが独自にこれらの決定を下す権限を与えられるべきなのでしょうか[64]。また、これらのシステム間の予期せぬ相互作用が、偶発的な分散型構造を介して問題を引き起こす可能性もあります。病室の音声アシスタントは、珍しい方言を持つ患者を一貫して誤認識する可能性があり[65]、診断推奨ソリューションのバイアスによって悪化し[66]、状況分析によって多くの誤った低レベルの要求では、スタッフはすぐに解雇され、スタッフは意思決定支援をオフにして、重大になる前に患者の根本的な問題を見逃すようになります[67]。

私たちのインテリジェントな病院はまた、既存の差別、不利益、プライバシーの懸念を増幅させる可能性があります。ソーシャルメディアやスマートフォンのデータから導き出された欠陥のあるAI行動プロファイリングは、たとえば、提供される治療法を決定する医療リスクプロファイルに影響を与える可能性があります。医療機器からのデータは、状況分析（血中酸素量、心拍数など）によってつなぎ合わされ、正確な予後を提供する可能性がありますが、それはスタッフによって暗黙のうちに蘇生しない（DNR）決定として扱われ、患者の最善の利益にはならないかもしれませんが、病院のリソースを最も効率的に使用する決定がなされます[68]。

AIを使用すると、病院では、センサーネットワーク（例えば防犯カメラ）で個人を特定し、プロファイル化した上で、個人でもグループでも人を識別し[69]、異なる扱いをすることが可能になります。この異なる扱いは病人にとっては恩恵となる可能性があります。病院は、患者の医療をニーズや好みに合わせて調整しながら、1日をスムーズにすることができるからです。しかし、時には有害である可能性もあります。これはまた、トイレ休憩時間を追跡したり、誰が誰と話しているかのマップを作成し、それを使用して組合を潰す目的で仕事とは無関係のグループを特定したり[70]、患者にどのような治療が提供されるかを決定したり、リソースが不足しているときにどの患者が治療されるかを決定したりすることで、不当なストレスを生み出す有害なものになる可能性もあるのです。このようなサービスは、業務システムに保存されている豊富な個人データ（キャプチャされたものと推測されたものの両方）に依存しており、インテリジェントな病院のシステムがハッキングされたり、個人データが漏洩したりするリスクを高めます。

受け入れから承認、信頼まで

企業が社会的営業免許を取得する方法を理解するには、この取り組みにおいて信頼が果たす主要な役割を検討することが重要です。運営するための社会的営業免許の利益は、地域社会がソリューション（この例ではインテリジェント病院）を受け入れ[71]と承認[72]した結果であり、この承認は、コミュニティに属する組織への信頼から生じています。インテリジェント病院の期待される利益を実現するためには、利用する地域社会の受け入れと承認を確固たるものにする必要があります。これを怠ると、混乱が生じ、コストが上昇し、メリットの実現が妨げられる可能性があります。このような混乱は、小規模なもの（床のセンサーを破壊したり、パターン化された服を使ってAIのプロファイリングや位置追跡を妨げたりするような小さな不服従行為）[73]から、大規模なもの（システムをハッキングして操作不能にしようとする試みや抗議行動）まで様々です。例えば、音声アシスタントにおける予期せぬバイアスは、恩恵を受けることのできないグループからの抗議につながる可能性があります。地域社会の規範や風土に沿っていない、または単にコミュニティの多くの人にとって疑問視される建設計画は、プロジェクト全体に失望される結果となる可能性さえあります。

重要なのは、どのような意思決定がなされているのか、どの意思決定が自動化されていてどの意思決定が自動化されていないのか、その意思決定が建物を利用する人々の仕事や私生活の質[74]にどのような影響を与えるのか、その意思決定が触れる人々の人間の尊厳に与える影響、そしてその意思決定が地域社会の期待とどのように一致するか、ということです。

インテリジェント病院を実現するためにAIをどのように使用するかについて、ある企業は大きな裁量を持っています。音声アシスタントには何らかの形の音声認識技術が必要になりますが、様々なオーディオおよびビデオ技術を使用して住民を追跡し、同様の効果を得ることができます。センサーの多くの構成、AI技術によってなされる可能性のある決定、（結果的にもたらされる）アクションなど、多くの異なるアプローチが可能であるが、病院で働いている人や病院を利用している人、そして病院を委託している会社の両方にとって受け入れられるものは限られており、望ましいものはさらに少ない場合もあります。

重要なのは、どのような意思決定がなされているのか、どの意思決定が自動化されていてどの意思決定が自動化されていないのか、その意思決定が建物を利用する人々の仕事や私生活の質[74]にどのような影響を与えるのか、その意思決定が触れる人々の人間の尊厳に与える影響、そしてその意思決定が地域社会の期待とどのように一致するか、ということです。企業はインテリジェント病院の社会的営業免許を必要としています。認可を得るためには、地域社会は企業を信頼する必要があります。すなわち、企業が表明する（そして実行している）ことを信頼し、コミットして実行する会社の能力を信頼することが必要です。

最終的には、信頼関係です。信頼とは、相手が特定の方法で行動するという信念であり、相手が信頼でき、有能であるという信念でもあります。地域社会と協力して作業し、ソリューションを形成し、オペレーショナルリスクを管理する方法について誠実さと能力を実証している企業は、ポジティブに受け入れられる可能性があります。地域社会の脆弱性を利用したり、皮肉屋や無能と見られたり、あるいは自らの脆弱性を十分に管理できていないと見られる企業は、悪い評価を受けることになるでしょう。

会社の制御が及ばない理由、または会社の労働の結果がコミュニティの期待と一致しないために、会社がコミュニティの期待に応えられなかった場合、信頼が失われます。信頼が崩れると、それはしばしば「地元のトラブルメーカー」からの「終わりのない要求」の結果、疑念に置き換えられます。

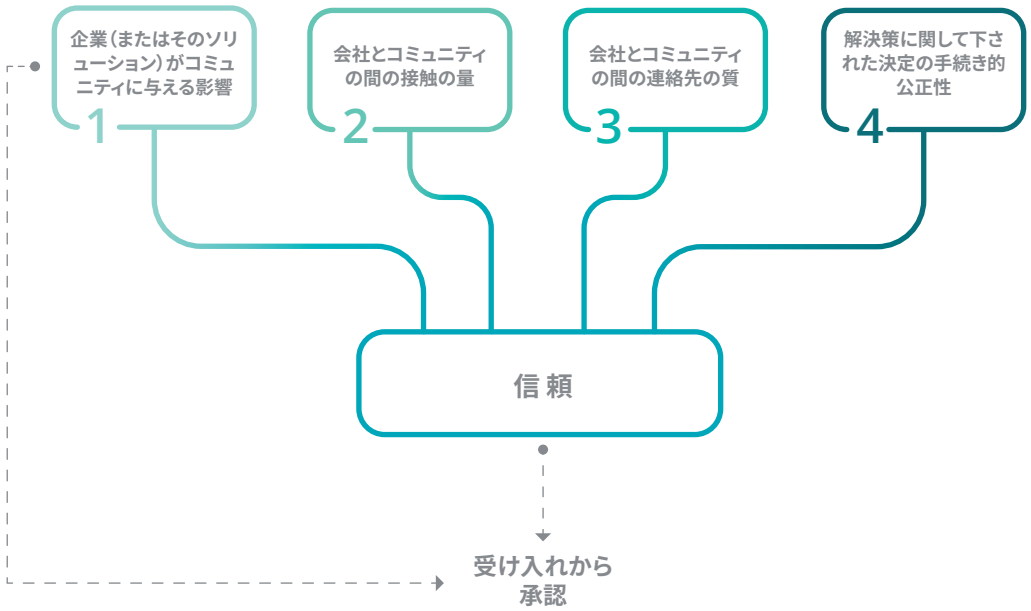
運営するための社会的営業免許のコンテキストにおいて、「信頼」は次の4つの要因に依存しています。企業（またはそのソリューション）がコミュニティに与える影響、会社とコミュニティの間のコンタクトの量。そのコンタクトの質。そして、解決策に関して下された決定の手続き的公正性（図1）[75]。企業は、これら4つの分野すべてで行動を起こし、地域社会との信頼を築き、地域社会からの行動の受け入れにより承認を高めることができます。

ソリューションが地域社会に与える影響を理解するには、すべてのソリューションにはメリットだけでなくデメリットも伴うことを認識する必要があります。インテリジェントな病院は、より効率的なエネルギー使用により、環境への影響が小さくなる可能性があります。スタッフがより幅広い言語に対応できるようにすることで、より包括的な運営を促進することができるかもしれません。また、診断がより正確で迅速になるかもしれません。一方で、建物にはストレスを増加させる可能性があり、機密性の高い個人データが漏洩したり、他の方法で悪用されたりするプライバシーのリスクを導入したり、望ましくない偏見、不平等、不利益を制度化したりする可能性があります。しかし、これらのメリットとデメリットの多くは企業が予測できるため、デメリットを軽減しながらメリットを強化することができます。

どのように地域社会がソリューション提供を体験するか、そして個人がどのように感じるのかを考慮することも重要です。例えば、Bluetooth対応の医療機器をインテリジェント病院のIoTネットワークに直接統合しても、先に説明したBluetooth対応の健康器具と同様の反応を受けるかもしれません。あるいは、画像認識ソリューションの周りでスタッフが協力する方法を単純化することで業務を合理化したいと考えても、プライバシーと人間の尊厳に関する懸念には十分に対応できないかもしれません[76]。このように、コ

図1

社会的ライセンスの信頼を取得するためには、4つの要因が関係します



Source: Adapted from Kieren Moffat and Airong Zhang, "The paths to social licence to operate: An integrative model explaining community acceptance of mining," Resources Policy 39 (March 2014): pp. 61–70.

ミュニティ内の異なる利害関係者が、ソリューションのメリットとデメリットに対して異なる期待を持つことは十分にあり得るのです。同様に、予期せぬ方言があった場合、音声認識の失敗に潔く対処しない限り、フラストレーションが発生したり、個人が排除されたりする可能性があります。このように、企業の意図と地域社会の期待との間にある解決策の影響や利益の不一致が、企業にとって予期せぬ結果をもたらす重要な原因となることがあります。

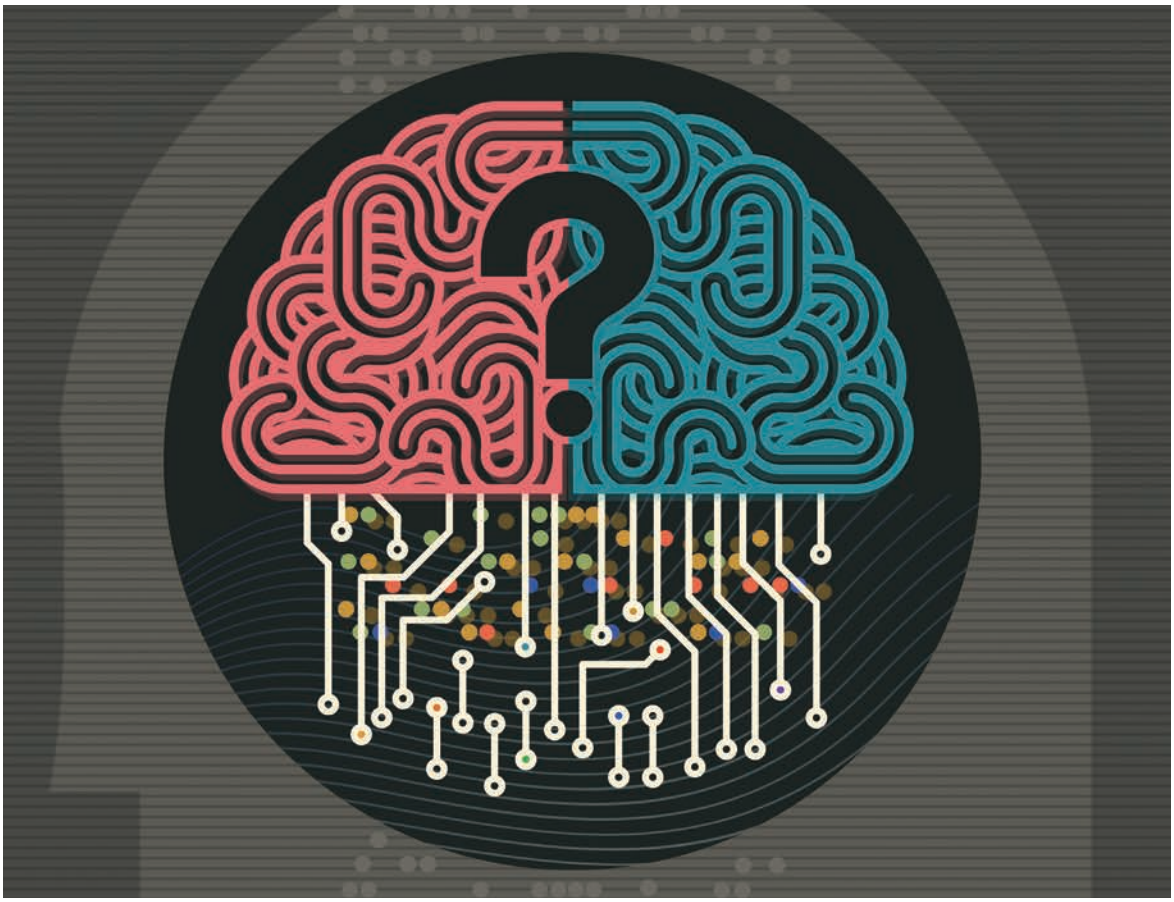
スマートな病院とインテリジェントな病院のインパクトの違い、AIのない病院とある病院の違いは、種類の違いではなく程度の違いです。AIは潜在的なメリットを増やしますが、リスクも高めます。

これにより、信頼をサポートする次の2つの要素、つまり企業とコミュニティとのコンタクトの量と質に続きます。信頼は、企業と地域社会の間の前向きなコンタクトが頻繁になされる結果です。病院を建てる企業は、地域社会に人間の顔、地域社会が信頼し、協力できる顔を提示する必要があります（結局のところ、企業は地域社会でもあります）。

コンタクトは頻繁（量）で、かつ意味のあるもの（質）でなければなりません。実際にはコンタクトは、ソリューションがコミュニティとそのコミュニティへの傾向にどのように影響するかを測定するための公式な影響調査から、コミュニティグループを介した、または個人と企業の代表者との間での日常的な連絡にまで及ぶ可能性があります[77]。ソリューションに直接影響されないが、結果に影響を与えることに関心のある利害関係者とのコンタクトも同様です[78]。このコンタクト一部は、一般データ保護規則（GDPR）などの規制や業界に関連する規制によって義務付けられる場合もあります。頻繁でかつ意味のあるコンタクトにより、コミュニティと会社はお互いについて学ぶことができ、誤解を最小限に抑え、自分の信念体系が他の人に投影されるのを避けることで、未知の（そして予期しない）ことを減らします。

信頼に影響を与える第四の要因である手続きの公正さは、ソリューションの開発と運用を支配する意思決定と紛争解決のプロセスです。個人は、意思決定プロセスにおいて合理的な発言権を持っていること、意思決定者が敬意を持って扱っていること、そしてその手順は公正であると見なしていることを認識しなければなりません。解決策が真実であるように、当事者（地域社会と企業）の間に平等な力関係があると感じる必要があります。

地域社会がインテリジェントな病院を受け入れ、その背後にある会社を信頼するためには、自分たちの意見が大切にされていること、自分たちの視点が説明されていること、自分たちが敬意・尊厳を持って扱われていること、自分たちの視点がソリューションに組み込まれていることを感じる必要があります。例えば、AIによって強化された終末期ケアや集中治療は、経済的な計算に基づくのではなく、患者をサポートし、尊厳と敬意を持って患者を治療するものです。例えば、病院全体で音声アシスタントを使用するという提案に対応し、問題が指摘され、代替案を提案する際には、個人やグループにとって実用的である必要があります。また、意思決定と紛争プロセスの両方とも、個人が自分の意見を受け入れ、地域社会の他の人の意見だけでなく、技術的、財政的な制約や会社自身の利益と照らし合わせて検討することができるように、個人が理解しやすく、比較検討できるようにする必要があります。



不足している議論について

運用におけるソーシャルライセンスの概念は、AIの道徳的ライセンスの強固な基盤を提供できますが、これからAIソリューションを開発する企業のニーズに適応させるための作業を行う必要があります。この記事でこれまで取り扱ってこなかった3つの質問があり、これらについて考えていかなくてはなりません。質問は次の通りです。

- (多くの利害関係者にとっては混乱を招く)不必要に技術的な詳細や、過度に抽象的な概念に戻ることなく、(提案された)ソリューションをどのように記述するのか?
- ソリューションの「コミュニティ」を構成するもの、つまり、利害関係者をどのように特定するか?
- 提案されたソリューションから、利害関係者が倫理的であると考えるソリューションに移行し、トレードオフをどこで行うか?また、そのトレードオフをどのように明らかにするのか?

これから順番に議論していきましょう。

ソリューションの説明

乗り越えなければならない最初のハードルは、例のスマート病院のようなソリューションを説明する方法を見つけることです。音声アシスタントに精通している場合は理解しやすくなりますが、状況認識と計画ソリューションを理解することは困難です。なぜならば、ソリューションは相互に関連する意思決定のネットワークにより病院内外からデータを取得し、様々な(潜在的な)患者の問題に対する推奨事項から、イベントを起こす要因となるからです。そのため、地域住民とそれを提案する人々がAIソリューションの形について話し合うために使用できる共通言語が必要です。共通言語があることにより円滑に、AIソリューションがどのような形になるのか、住民の位置情報を追跡し、追跡データをどのように使用するのか、状況認識とどのように相互作用するのか、状況認識がどのようなアクションやプロセスを駆動するのか、などを議論ができるのです。そして、幅広いセンサーとエフェクターを統合することで、分離された自動決定を含むスマートホスピタルを、統合された自動決定ネットワークを含むインテリジェントな病院に変換するAIに対する機能のニーズが浮き彫りにします。その結果、状況認識が推進できるプロセスなど、機能を実現するための代替アプローチやメリットとデメリットの比較検討の議論ができるのです。

ソリューションを説明すると、「醸造問題」と呼ばれるものを解決することになります。

ソリューションを説明すると、「醸造問題」と呼ばれるものを解決することになります。醸造が工芸から工学へと移行する前に、生物学と化学を統合した言語である微生物学を開発する必要がありました。これにより、醸造プロセスを微調整し、より一貫した結果を得ることが可能になりました。同様に、もしAIソリューションを微調整する場合、コミュニティとそれを提案する人々の両方がアクセスできる言語、つまり倫理と実装の両方を網羅しながらも、技術的な詳細をあまり入れずに、それを記述し、議論できる必要があります。理解しやすく、かつ有用であるためには、この言語は、高レベルの倫理原則よりも具体的である必要がありますが、実施の詳細よりも一般的な内容である必要があります[79]。また、技術的な専門用語は避け、信頼の構築に貢献する共通の理解を支えるために、わかりやすく親しみやすい用語を使用すべきです。提案するソリューションの中で、相互に関連して集約された一連の意思決定(意思決定とその関係)を記述できるようにする必要があります。どのアクター(人間または機械)が各決定を実行するか、どの情報が決定を左右するか、決定から生じる結果(および情報)、そして、これらの行動(および情報の変化)が人間に与える影響[80]を記述します。

人間と機械は異なる方法で考える(決定する)ので、人間によってなされた決定とAIによってなされた決定を区別することは重要です[81]。人間として、私たちは意思決定をするときには、無意識のうちに行っていることであっても、感覚と生活経験を利用します。私たちは異常で予期しないことに気づき、それを私たちの審議に織り込みます。一方、機械は、考慮するように設計されたデータのみを考慮します[82]。ミサイルを発射する、個人の社会的利益を撤回する、または救命機械を別の患者に移動するなどの決定が結果的である場合、決定は人間によって下されるのが一般的です[83]。それは、人間は異常な要因を考慮できるからであり、これは予期しないことを意思決定するためには大変重要なことです。あるケースでは、規制によって、アルゴリズムではなく人間(またはグループ)による特定の決定が必要になることがあります[84]。しかしながら、私たちは人間と機械の決定を区別したい一方で、機械の決定がどのように実装されるかにはあまり興味がないこともあります。

インテリジェントな病院は、どのような情報が収集されるか、この情報によって通知される決定、決定を行う主体(人間または機械)、および各決定から生じる情報とアクションの観点から説明されることがあります。例えば、ある説明においては、来訪者に発行された一時的な識別バッジ(情報)をビデオ画像と声紋(情報)に関連付けて、訪問者を(機械の決定によって)識別、病院が来訪者を追跡しながら建物の中を移動します(この時、2つを関連付けるために使用される技術に関する情報は、関連付けが行われるという事実よりも重要ではありません)。訪問者が禁止区域に迷い込んだと判断した場合、機械による判断でフロアのセキュリティスタッフに通知します。ソリューションの完全な説明には、これらの情報-決定-行動の流れの多くが含まれており、私たちのインテリジェントな病院の操作をカバーし、コミュニティとのコラボレーションで進化し、洗練されていきます[85]。

コミュニティの定義

作業を始める前に、私たちはシステムの社会的境界を明確にする必要があります。私たちは、利害関係者が誰であるかを定義し、彼らの立場を理解し、活動の社会的理念を発見し、専門家や情報提供者などを特定しなければなりません[86]。

「コミュニティ」あるいは「地域社会」という用語は狭すぎるため、インテリジェントな病院のように、複雑なソリューションが影響を及ぼし、その生活に影響を与えるケースにおいては利害関係者の多様性を捉えることができません。SLOは、明確に定義されたコミュニティによって付与された単一のライセンスであると簡単に想定できます。しかし、コミュニティが他の地理的地域やコミュニティから集められた多様なサブグループの集合体で構成されている複雑な環境では、そう簡単ではありません。このような場合には、SLOをこれらサブグループにまたがる複数のライセンスの連続体として、重なり合い、相互に関連する複数のコミュニティにまたがって機能すると考える方がより生産的です[87]。

人類学者は、考慮すべきさまざまな行動、考え、態度を、雇用状況、収入、性別、第一言語などの人口統計学的属性（コミュニティの違いを説明する要因）とともにリストアップすることから始める場合があります。これらの要素は一連のコミュニティ要素[88]にマッピングされ、各要素はコミュニティに存在する可能性のある緊張や好みの違いを捉えます。インテリジェント病院の例は、性別に対する労働者の態度（性別が厳密に二元的であると見なされるか、より広い定義が受け入れられるかどうか）、彼らの仕事の性質（分析的で官僚的なものか手作業であるか）、教育を受けているか、宗教や信念体系、社会経済的な（不利な）優位性、あるいは病院で定期的に働いているか、たまにしか来院しないかなどです。これらの要因により、私たちのコミュニティがカバーする可能性のあるマップ[89]を描くことができる。

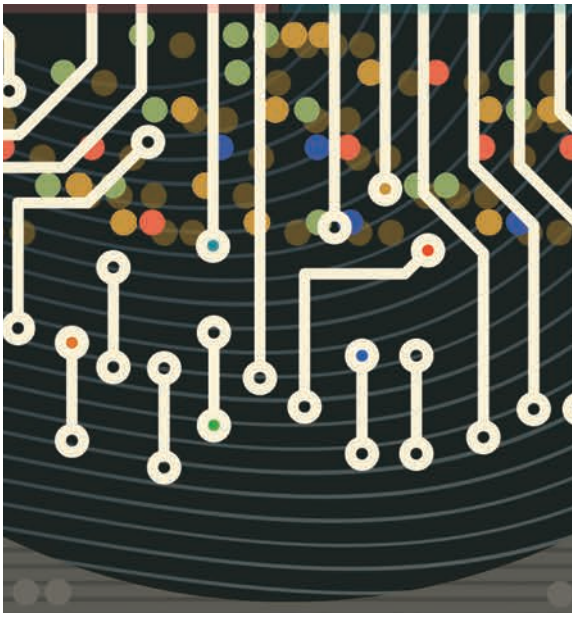
企業は、観察、構造化および半構造化インタビュー、グループディスカッション、日記調査、調査対象のグループメンバーとのワークショップなど、コミュニティメンバーの行動、考え、態度を調査するために、様々なフォーマルな方法とインフォーマルな方法を利用します。重要なのは、研究者と被験者の間で情報が行き交うオープンな対話を確立することです。地域社会から参加者を選択して、特別な問題を含む可能性がある状況に特に注意を払いながら、すべての既知の要素を確実にカバーすることができます。目標は、地域の歴史とその中の個人について可能な限り多くのことを学び、地域社会が含む社会的世界とその機能についての完全な理解を深めることです。

企業が学習したことは、地域社会の代表的なメンバーのプロファイル（および代表的なメンバー）を特定するアクターネットワーク[90]に取り込まれ、これにより提案された解決策がどのように関連しているかを把握することに活用できます。なお、アクターネットワークは、人間同士の繋がり・関係性を可視化することができ、その関係、対立、同盟、そしてそれらを結びつけるプロセスを記述するものです。

ソリューションの改良

私たちの最後の課題は、コミュニティと協力してソリューションを洗練させることです。一般的形態素解析[91]（GMA）の技術に触発されたアプローチでは、これを4つのフェーズに分けることができます。

まず、インテリジェント病院などのアイデアを挙げ、その説明を作成します。建物はこのデータを使用してこれらの決定を行い、この決定がこれらのアクションをもたらすことになります。これは、先ほどの記事で説明した言語であり、建物が建物内でどのように訪問者を監視し、診断をサポートし、緊急事態を特定して管理を支援するかを説明する情報決定アクションです。この時点では、例えば、意思決定が機械によって行われるのか、人間によって行われるのかにはこだわらないことで、一般的な説明にとどめておくことができます。



次に、2つの段階を経て、ソリューションを洗練させていきます。技術的に不可能なものを排除すること、そしてコミュニティにとって許容できる（そして受け入れられる）ものを発見することです（通常、社会規範は進化し続け、規制は遅れをとっているためです）。

不可能な要素を排除するためには、可能なソリューション構成（どの情報の組み合わせが、どのアクションをトリガーする決定をもたらす可能性があるか）を全て列挙し、技術的に不可能な構成や規制によって禁止されている構成などを検討する必要があります。

不可能な要素を排除するためには、可能なソリューション構成[92]（どの情報の組み合わせが、どのアクションをトリガーする決定をもたらす可能性があるか）をすべて列挙し、技術的に不可能な構成や規制によって禁止されている構成など検討することとする必要があります。明らかに不可能なものを排除することを意味します。規制により、特定の決定が人間によってなされなければならない、または人間によって監督されなければならないことが要求される場合があり、その場合、ソリューションの説明に「この決定は人間によって実行されます」と追記します。たとえば、インテリジェントな病院では、救命機器を優先度の高い患者に移すという決定は人間が行う必要があるでしょう。

人間と機械による意思決定の両方の利点が欲しい場合には、意思決定を2つに分けるかもしれません。人間と機械の両方の意思決定のメリットが必要な場合は、意思決定を2つに分割できます。人間の意思決定の一部と見なすことができる機械の提案です。状況分析と計画ソリューションは、行動方針の意思決定する責任のある人間の管理者に限定されるかもしれません。あるいは、適切な資格を持った人や、医療専門家のような特定のレベルの年功序列を持った人が決定を下す必要がある場合があります。また、機械の決定が人間にも理解できることを必要とする場合もあります。使用される技術がどのようなものであっても、それが行う決定の根拠を提供しなければならないことに注意してください。例えば、計画エンジンは、機械学習よりもルールベース[93]を通して実装する方が適切な場合があります。これにより、ユーザーがソリューションの推論を操作したり調整したりすることが簡単になります。

分析における最初のフェーズでは、データの一部が個人データ（性別など）を表していて、少数の特定の意思決定のための入力としてしか使用できない場合も判断します。この分析に基づいて、要素（情報、決定、アクションとそれらの関係）を変更するか、各要素の使用方法または実装方法を制限するようにアノテーションを付けることによってソリューションの説明を進化させることができます。

次のステップである許容できないものを削除することも同様のプロセスですが、コミュニティと協議して行う必要があります。コミュニティの代表者（以前に特定された代表的なコミュニティメンバーのプロファイルに合わせて）と協力することで、企業は、どのような結果やプロセスがコミュニティに受け入れられやすいか、あるいは受け入れられにくいかを特定することができます。

顧客またはユーザーにサービスを提供するために必要なコンポーネントをマッピングした、ビジネスまたはサービスの構造をマップしたWardley map[94]などのツールを使用して、ソリューションの（コミュニティにとっての）メリットと成

熟度を調査し、仮定を明らかにし、チャレンジを許可し、コンセンサスを形成することもできます。例えば、ある特定の決定が「公正」であることが求められる場合、例えばCOMPASにおける平等と公平の選択、または緊急時の患者ニーズの優先順位付けなど、どのように公正さを実現するかは、コミュニティの代表者と協力して決定し、決定の説明に記載することができます。また、音声やタッチのインターフェースと画像認識を統合した自動登録プロセスなど、関連するコンポーネントのグループは、特定のAIコンポーネントが完璧ではなくても、全体として不利益やその他の悪影響を与えないように見直すことができます。また、議論を呼ぶ可能性のあるテクノロジーに対するコミュニティの態度も考慮する必要があります。コミュニティは、ユビキタスなビデオ監視に不快感を示すかもしれないので、このインテリジェントな病院のオーナーは、建物内を移動する住民を追跡するための、より受け入れやすい方法を見つけるように求められます[95]。このような難しい問題の場合、地域全体に受け入れられる首尾一貫したアプローチを開発するために、地域の様々なグループと相談する必要があるかもしれません。

不可能を排除し、何が許容されるかを発見した時点で、ソリューションの詳細な概要を示しますが、企業や地域社会が適切と見なせる詳細事項が通常ありません。そのため、完全なソリューションではありません。例えば、建物ゾーンの温度を維持するために使用されるアルゴリズムは、指定されないままになる可能性があります。一方、その他の詳細は、防犯カメラから発せられるビデオストリームの許容される用途、AIソリューションからの結果的な勧告（正確な予後など）をどのように扱うべきか、行動プロファイルが意思決定に影響を与える範囲、どの機械の意思決定が人間に理解可能であることが必要か、相反する患者の優先順位に対処する際の「公正」の解釈方法など、非常に厳密に指定される場合があります。

不可能を排除し、何が許容できるかを発見するプロセスにより、企業は地域社会と協力して、倫理原則（公平性や危害の防止など）がどのように制定されているかを判断し、ソリューションの共有説明として文書化できます。これは「倫理的要件アーキテクチャ」と呼びます。

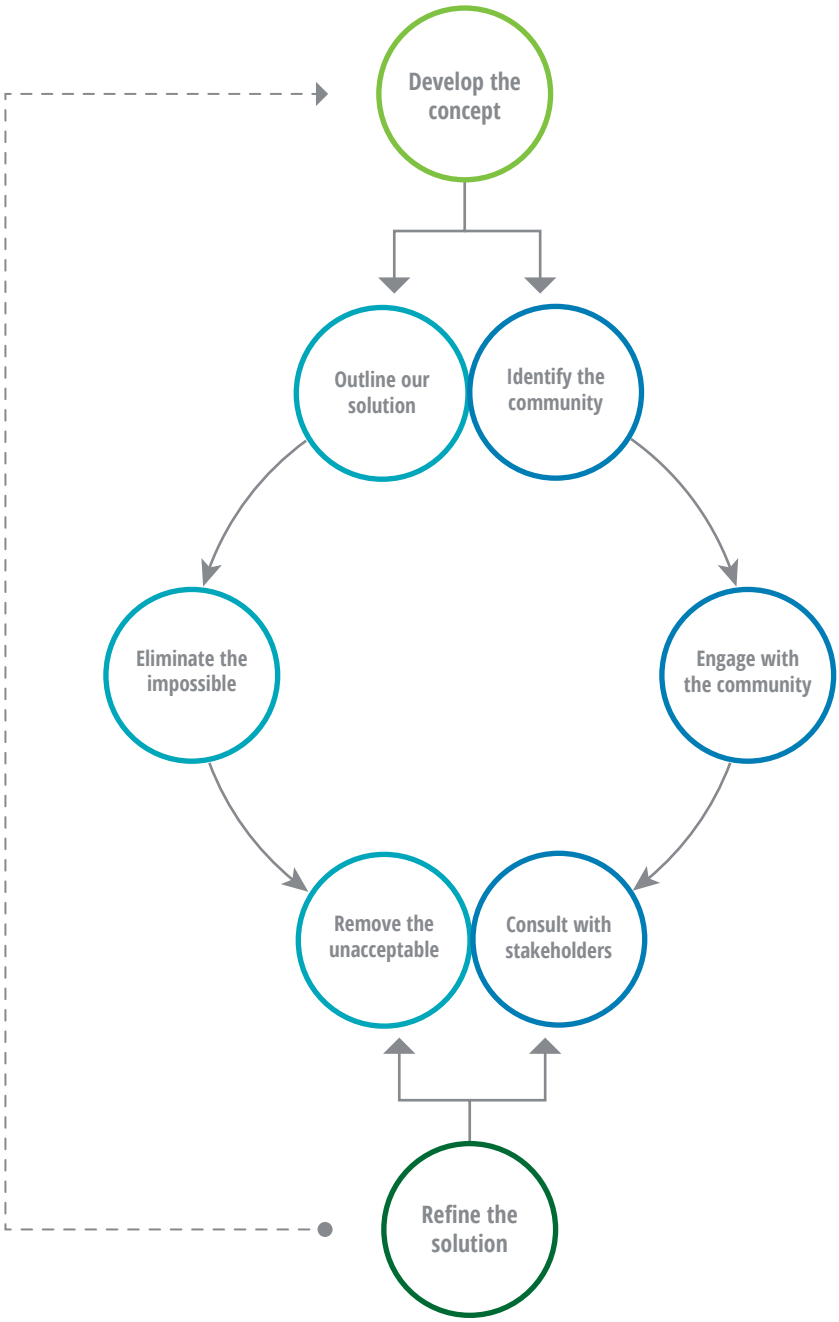
最後の第4段階は、洗練されたソリューション記述をどのように実現すべきかを決定するという技術的な課題です[96]。この段階では、偏りのない倫理的なアルゴリズム、つまり「信頼できるAI」[97]を作成するための方法論と技術に関する研究が活用します。

倫理的要件アーキテクチャの開発

これらのフレームワーク（社会的ライセンス、倫理的要件のための言語、社会科学的アプローチによるコミュニティの理解、GMAによるソリューションの開発と改良）を統合すると、図2に示すようなものができるかもしれません（ただし、この点はプロセスの概要や説明を開発する目的で十分なものに過ぎず、プロセスそのものを開発するものではないことに注意ください）。

図2:

利害関係者と人工知能ソリューションを検証および改良するためのフレームワーク



Source: Deloitte analysis.

図2の円の左半分は、最初のソリューション提案から改良を経て、倫理的なソリューションアーキテクチャと最終的なシステムを構築する、GMAのプロセスを表しています。円の右半分は、社会科学の流れを表しており、ここでは、企業がコミュニティの要因を特定することでコミュニティ図を作成した後、影響分析によってコミュニティの関係図をマッピングし、その後、ソリューションを洗練させるためにコミュニティと関わります。

これらの流れの間には2つのタッチポイントがあります。最初の段階では、企業は最初のソリューションの概要と、コミュニティの構成要素に対する理解を並行して進め、2番目の段階では、企業がコミュニティの利害関係者と協力して、受け入れられるソリューションがどのようなものかを形成していきます。このようにして倫理的なソリューションのアーキテクチャを作成することで、例えば、インテリジェント・ホスピタルの利用者が監視への懸念からソリューションのコンセプトそのものに抵抗したように、コミュニティがそもそもソリューションを求めていない可能性に対応し、企業の動機や目標を透明にすることでコミュニティの信頼を確立することができます。また、この記事の冒頭で指摘したように、SLOはソリューションやコミュニティの進化に合わせて見直し、維持する必要があるため、プロセス全体もフィードバックします。

人間が「良い」と考えるものは、人間が集団で共存し、最終的には調和のとれた文明を築くために進化してきました。モラルには、「殺してはいけない」「盗んではいけない」など、何千年も前から存在するものもあれば、「包括的投票権」や「結婚権」など、比較的最近になって登場したものもあります。また、特定のコミュニティに固有のモラルも存在します。私たちが共感すべきなのは、コミュニティの集合体が相互に影響し合うことで生まれてくるモラルの質なのです。AIソリューションがコミュニティに受け入れられ、信頼されるためには、AIソリューションが適用されるコミュニティと、そのコミュニティに生まれつつあるモラルの両方に関わる必要があります。その関わりは意味のあるものでなければならず、継続的でなければならず、コミュニティが道徳的で「正しい」と信じるものがソリューションに組み込まれていなければなりません。

ここで重要なのは、この議論はオープンマインドで行うべきだということです。閉鎖的で共感できない、あるいは単に関心がないという考え方では、正反対の結果をもたらすでしょう。適切なスキルとマインドセットがなければ、このプロセスは望ましい結果をもたらさないかもしれません。企業とコミュニティの人々との交流は、相互に受け入れ可能な結果に向けて、信頼と尊敬に満ちた関係を築き上げるためのタッチポイントです。先に述べたように、課題はどちらか一方を優遇するのではなく、社会の世界を橋渡しすることです。

また、企業が誠意をもってコミュニティに関与しないことを選択する可能性や、AIを開発する企業とその影響を受けるコミュニティとの間に力の非対称性がある場合、最悪、企業の意図が悪意に満ちたものである場合なども考慮する必要があります。このようなケースにおいて、AIのモラルライセンスのようなフレームワークがあれば、規制当局は、悪質な企業が少ないとしても、既知の方法を用いて、コミュニティとのやりとりとその結果を文書化していることを確認するための報告義務を制定するための基盤を得ることができます。倫理的要件アーキテクチャ」の開発は、AIの環境影響調査に相当する規制となる可能性があります。

AIの道徳的ライセンスの必要性

倫理的なAIに関する研究は、倫理的なAIを実現するために必要な原則、要件、技術基準、およびベストプラクティスの策定に焦点を当てています。しかし、AIは倫理的であるべきだという明確なコンセンサスがあり、倫理的なAIの原則については世界的に収束しているものの、これらの原則をどのように実現すべきか、「倫理的なAI」が実際に何を意味するのかについては、本質的な相違が残っています[98]。

本稿は、「倫理的なAI」に関するものですが、倫理とAIの問題を直接扱うことはなく、別のアプローチをとっています。どのようなAIの使い方が倫理的であり、そうでないかを定義しようとするのではなく、企業が関わるコミュニティと協力して、運用したいAI対応ソリューションの道徳的ライセンスを取得し、維持する必要があると提案しています。さらに、企業は、現在AI技術と考えられているものを含むソリューションだけでなく、意思決定を自動化し、他の業務システムと統合して意思決定ネットワークを構築するあらゆるソリューションに対して、このような取り組みを検討する必要があります。

このようなアプローチの違いは、次の3つの見解によるものです。

- AIソリューションは、「公正」や「倫理的」なアルゴリズムや開発手法を開発しても、倫理的なものにはならないこと。

- あるソリューションが倫理的であるかどうかを判断できる単一の世俗的な社会（完全に正常化された社会世界、客観的な基準）は存在しないこと。

- 倫理的なAIの重要性は、破壊的なAI技術の開発や、独立した自己認識型AIソリューションによる実存的な脅威によるものではなく、むしろ自動化された意思決定ネットワークの広範な出現によるものであること。

倫理的なAIとは、特定の技術やソリューションの偏りや失敗を管理するための規制、技術、方法論の開発であり、それだけでは十分ではありません。倫理とは、そこに到達するためのルール、行動、振る舞いのことです。私たちが目指すべきは、モラルのあるAIです。そのためには、目的と手段を明確にしておく必要があります。多様で開かれた社会では、何かをすべきかどうかを判断する唯一の方法は、私たちの行動によって影響を受けるコミュニティとオープンに協力し、信頼を得て、提案を受け入れてもらうことです。

文末脚注

1. Luc Zandvliet and Mary Anderson, “Introduction,” Getting it Right: Making Corporate-Community Relations Work (Sheffield, UK: Greenleaf Publishing Limited, 2009), p. 5.

2. ABC News, “Bentley gas protest makes history,” May 20, 2014.

3. ABC News, “NSW government buys back Lismore CSG licence for $1 million,” October 19, 2015.

4. ここでは、「モラル」を私たちが求める結果であり、「倫理」はそこに到達するためのプロセスであるとしています。「何が、自分たちまたは会社にとって大切なのか」「正しいこととは何か」この価値観に基づく、考え方、法律、社会のルールの融合が「エシックス」です。

5. プロジェクトの損益計算書に直接計上されるコストもあれば、風評被害のように定量化が難しいコストもあります。

6. 社会的営業免許（Social License）の運用についての概要は、Joel Gehman, Lianne M. Lefsrud, and Stewart Fast, “Social license to operate: Legitimacy by another name?,” Canadian Public Administration 60, no. 2 (June 2017): pp. 293–317をご参照ください.

7. my.Flow, accessed July 1, 2020; Jordan White, “Bluetooth tampons—YES!,” SmartFem, 2016; Ashley Carman, “A Bluetooth-connected tampon. Hoo boy.,” Verge, May 18, 2016; Gemma Mullin, “Controversial bluetooth tampon lets you know when it needs changing—but has 12ins string,” Sun, January 8, 2020.

8. Tim Brennan and William Dieterich, “Correctional Offender Management Profiles for Alternative Sanctions (COMPAS),” in J.P. Singh et al. (eds.), Handbook of Recidivism Risk/Needs Assessment Tools (Chichester, UK: John Wiley & Sons, Ltd, 2018), pp. 49–75.

9. Julia Angwin et al., “Machine bias,” ProPublica, May 23, 2016.

10. Stephanie Wykstra, “Government’s use of algorithm serves up false fraud charges,” Undark, June 1, 2020.

11. Kranzbergの第一法則として知られる。From Melvin Kranzberg, “Technology and history: ‘Kranzberg’s Laws,’” Technology and Culture 27, no. 3 (July 1986): pp. 544–60.

12. B. J. Fogg, Persuasive Technology: Using Computers to Change What We Think and Do (Burlington, MA: Morgan Kaufmann Publishers, 2002), p.

13. Ben Loewenstein, “Regulation of AI: Not if but when and how,” RSA, November 21, 2017.

14. 業界が規制を求めるのは、ある意味で危険なことです。企業は通常、大きなリスクがない限り規制を求めないからです。最近、多くの企業が顔認証技術の開発中止を決定したのは、規制ができないことが一因と考えられます。詳しくは、Bobby Allyn, “IBM abandons facial recognition products, condemns racially biased surveillance,” NPR, June 9, 2020.を参照ください。

15. Brent Mittelstadt, “Principles alone cannot guarantee ethical AI,” Nature Machine Intelligence, November 2019.

16. また、データプライバシー（ISO29100:2011）のような複雑な規制の枠組みの中で原則を実施することは困難な場合があることに留意してください。

17. 「社会的世界」とは、社会学で頻繁に使われる用語で、共通のシンボル、組織、活動が生まれる「言説の世界」を指します。社会的世界は、物理的な境界線を必要としない文化的な領域を含みます。政治、科学などの「社会的世界」がその典型例です。

18. Nils J. Nilsson, “Preface,” The Quest for Artificial Intelligence: A History of Ideas and Achievements (Cambridge, UK: Cambridge University Press, 2009), p. 13.

19. Bowerbird (バウワーバード)はオーストラリアに生息する鳥で、オスは精巧な構造物（バウワー）を作る習性がある。羽根や貝殻などの色鮮やかな装飾品で飾られた東屋は、求愛のために雌を惹きつけます。

20. 例えば、データサイエンスとAIの境界線は曖昧で、この2つの分野は多くの技術を借りています。

21. Marvin Minsky氏などの言い換え。

22. Rodney Brooks氏は、オーストラリアのロボット工学者であり、オーストラリア科学アカデミーのフェロー、作家、ロボット工学の起業家です。ロボット工学における行動主義的アプローチを広めたことで知られています。

23. Jennifer Kahn, “It’s alive!,” Wired, March 1, 2002.

24. 自律走行車のように、数年前にはサイエンスフィクションのように思えたソリューションを、既存の技術を使ってかなり簡単に開発できるようになったことからもわかりますが、現在の技術水準を超えて進化させることは非常に困難です。これらのソリューションは指数関数的に向上するのではなく、コストと労力が指数関数的に増加し、性能はわずかにしか向上しないという状況です。
Stefan Seltz-Axmacher, "The end of Starsky Robotics," Starsky Robotics 10-4 Labs, March 19, 2020をご参照ください。

25. Matthew Hutson, “Core progress in AI has stalled in some fields,” Sciencemag.org, May 29, 2020; Economist, “An understanding of AI’s limitations is starting to sink in,” June 11, 2020.

26. CRISPRは遺伝子編集技術で、病気や害虫に強い作物の開発から、最近ではCOVID-19の原因となるウイルスの診断テストを可能にするなど、農業や公衆衛生の分野で幅広く活用されています。

27. Kranzberg, “Technology and history: ‘Kranzberg’s Laws.’”

28. 著者らは以前、この変化と、それが私たちとテクノロジーとの関係をどのように変えているかについて述べています。
Peter Evans-Greenwood, Robert Hillard, and Alan Marshall, The new division of labor: On our evolving relationship with technology, Deloitte Insights, April 9, 2019をご参照ください。

29. 例えば、都市では、インフラや交通ネットワーク、サービス提供を管理するシステムに自動化された判断を統合しています。バスの時刻表やメンテナンスのスケジュールから、取り締まりのパターンまで、あらゆるものが "スマート"になっています。詳しくは、Beryl Lipton, “Smarter government or data-driven disaster: The algorithms helping control local communities,” MuckRock, February 6, 2020. をご参照ください。

30. ここでは、より一般的な用語である「サイバー・フィジカル・システム」ではなく、「ディシジョン・ネットワーク」を使用しています。これは、不道徳な自動化された意思決定が、統合された意思決定のネットワークにどのような影響を与えるか、また、それに伴うモラルハザードに焦点を当てているためです。

31. IoT（Internet of Things）とは、1994年頃に作られた造語で、インターネットに接続された小型でスマートな機器のネットワークを意味しています。“bsy's list of internet accessible coke machines,” August 29, 2003.をご参照ください。

32. エフェクターとは、生物学の用語で、刺激に反応して行動する器官や細胞のことです。ロボット開発者の間では、世界に影響を与えるために使用できる車輪や腕などの付属物を指す言葉として採用されています。別の用語として「インフルエンサー」があり、これは一部の軍事ドクトリンで使用されており、「センサー」と「シューター」のAI分類を補完するものですが、これは一般的に使用されている用語とは一致しません。

33. アルゴリズムによるモラルハザードとも言えます。

34. 事故を起こす可能性があるため、動いている車を停止したくありません。

35. 例として、Greg Jennett, "Robodebt removed humans from Human Services, and the government is facing the consequences," ABC News, May 29, 2020; Wykstra, "Government's use of algorithm serves up false fraud charges." を参照ください。

36. Ibrahim Diallo, “The machine fired me,” iD, June 17, 2018.

37. John Thorpe, The Information Paradox: Realizing the Benefits of Information Technology (McGraw-Hill Higher Education, 2003).

38. これまでにも、さまざまな哲学的難問を解決するための思考実験には多大な努力が払われ、問題の範囲も狭められてきました。しかし、このような考え方を現実の世界に適用することは明確ではありません。例えば、自律走行車が、人口統計学的または身体的特徴によって2人の人間を識別し、どちらにぶつかるべきかを倫理的原則に基づいて判断しなければならないという問題が考えられるかもしれません。しかし、既存の技術や想像上の技術では、このような識別を十分に行うことができないため、これは無意味であると思われます。この問題の微妙なバリエーションをすべて説明することはできませんし、原理間の対立をすべて予想して解決することもできません。ブレーキをかける、障害物がないときだけハンドルを切る、その他の問題は保険や賠償責任で処理する、といった小さなシンプルな原則が使われる可能性が高いのではないでしょうか。詳しくは、James Wilson, "The trolley problem," Aeon, May 28, 2020. をご参照ください。

39. Philippa Foot, “The problem of abortion and the doctrine of the double effect” in Philippa Foot, Virtues and Vices and Other Essays in Moral Philosophy (New York: Clarendon Press and Oxford, UK: Oxford University Press, 2002).

40. トロッコ問題を自律走行車に適用する際に問題となるのは、この思考実験が当事者でない人間を中心に設計されていることです(対象者は、その行為から離れていたり抽象化されていたりすると、集団よりも個人を殺す可能性が高いが、被害者に触れなければならないときには躊躇したり拒否したりする)。この（インパクトへの）近接性の要因は、この論文で採用されている多元的なアプローチや、逆にモラルハザードの問題においても重要です。

41. Self-driving cars are a case in point. See Kelsey Piper, “It’s 2020. Where are our self-driving cars?,” Vox, February 28, 2020.

42. Naaman Zhou, “Volvo admits its self-driving cars are confused by kangaroos,” Guardian, June 30, 2017.

43. Will Douglas Heaven, “Google’s medical AI was super accurate in a lab. Real life was a different story.,” MIT Technology Review, April 27, 2020.

44. Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, “On the (im)possibility of fairness,” arXiv:1609.07236v1 [cs.CY], September 23, 2016.

45. Gerald Gaus, The Tyranny of the Ideal: Justice in a Diverse Society (Princeton, NJ: Princeton University Press, 2016)の162頁を参照ください。"私たちが生きている社会世界の性質についての意見の相違は、周辺的なものでもなければ、価値や優先的な紛争として再記述できるものでもない（つまり、視点の評価基準の要素に押し込められる）ことを強調しておきます。私たちの最も深く、最も困難な紛争のいくつかは、価値や正義の原則についてではなく、これらの原則が適用される世界についてのものです。最もわかりやすい例は、妊娠中絶の権利をめぐる長年にわたる根強い争いです。中絶の権利を擁護する人たちは、この事件は個人の自律性に関する基本的な権利に関わる決定的なものであると考え、中絶の権利に反対する人たちは、女性が自分の体をコントロールするという主張にあまり敏感ではないと描かれています。中絶に反対する人たちは、そのような自律性に深く傾倒することはあっても、それが他の人の生きる権利を覆すことになるような場合には、そうではありません。もちろん、抽象的には、中絶の権利を擁護するほとんどの人も、そのような状況では引くでしょう。この論争は、自律性と生命の権利の原則が適用される社会的世界についての中心的な問題です。この2つの社会世界には同じ人間が存在しないので、抽象的な正義の原則について完全に合意しても、この論争は解決しません。多くの道徳哲学者が、完全に正常化された単一の世俗的な社会世界しか存在しないという点で同意しているからこそ、この論争は単に価値観や抽象的な正義の原則に関するものと誤認されなければならないのです。

46. Steve Jacobs, “How views on ‘when life begins’ drive Americans’ abortion attitudes,” Heterodox Academy, July 9, 2020.

47. 私たち全員が住んでいる「社会」は制度の基盤であり、すべての規制や規範が、すべての視点に受け入れられる共有された共通の道徳的存在に照らして測定される、正常化された社会世界です。倫理や正義について推論することができる唯一の正規化された視点は存在しません。完全に公正な社会の独自の公平な解決という特殊な問題の核心は、複数の競合する正義の理由が持続可能であることであり、これらの理由はすべて公平性を主張しているが、それにもかかわらず、互いに異なり、競合します。詳しくは、Amartya Sen, “Introduction” in The Idea of Justice (Cambridge, MA: Belknap Press of Harvard University Press, 2011), p. 12.を参照ください。

48. 「正義」などに関する私たちの仮定が、私たち自身の生きた経験に固有のものであることを示す良い例が、Sam Dubal, “Against humanity: What the Lord’s Resistance Army can teach us about flaws in the ideal of human rights and the fight for justice,” Aeon, March 18, 2020.に示されています。

49. 功利主義は、影響を受ける個人の幸福と福利を最大化する行動を促進させるものです。

50. つまり、平等ではなく、公平性を優先した社会世界です。

51. Angwin et al., “Machine bias.”

52. Tafari Mbadiwe, “Algorithmic injustice,” The New Atlantis 54 (Winter 2018): pp. 3–28.

53. 著者らは、倫理的なAIのための健全な原則として、以下の論文を使用しています：Jim Guszcza et al.、Human values in the loop: Design principles for ethical AI, Deloitte Insights, January 28, 2020.

54. これは、Thorpe氏による「4つの領域」や、禅の修行法である "誰が質問しているのか？"を問うことが良い補足となるでしょう。「何が倫理的であるかを決めるのは誰か」という問いに答えようとするならば、質問に答えるグループと質問の答えによって影響を受ける人々との関係、そして質問に答えるグループの意図を4つの領域の観点から理解する必要があります。投げかけられた質問に答えるとき、誰が含まれ（誰が除外され）、どのような要素が考慮されたのかです。

55. C. P. Snow氏は、熱力学の法則を次のように覚える方法を持っていたと言われています。
-ゲームをしなければならない（物理的世界は避けて通れないものであるため）
-勝てない（つまり、物質とエネルギーは保存されるので、何かを無駄に得ることはできない）
-損益分岐はできない（常に無秩序が増加するため、同じエネルギー状態に戻ることはできない）
-ゲームから抜け出すことはできない（絶対零度は達成できないため）
Wikiquote, "Thermodynamics," accessed July 1, 2020.

56. 実際、私たちの周りにはすでにこのような自動意思決定ネットワークが存在している可能性が高く、そのデメリットはまだ重大な問題として認識されていないのです。

57. SLOを得るためには、（社会的規範による）拒絶、受容、そして最後に（信頼のために採用したバイアスによる）承認という3つの閾値を通過しなければなりません。詳細は、Robert G. Boutilier and Ian Thomson, “Modelling and measuring the social license to operate: Fruits of a dialogue between theory and practice,” Shinglespit Consultants Inc., 2011.を参照ください。

58. 「コミュニティ」という言葉は、複雑なソリューションが関わり、重なり合い、相互に関連する社会集団の網を表現するには、狭すぎる言葉であることが多いことに注意してください。この点については、後で詳しく説明します。

59. 「倫理的なルールネットワークのための方法」は「倫理的なニューラル・ネットワーク分類器のための方法」とは必然的に異なるものとなります。

60. これは、特定のソリューションが倫理的であるかどうかの問題は、第三者やその他の外部機関ではなく、そのソリューションの影響を受けるコミュニティの問題であることを意味します。ただし、国内の規範や規制が適用されるため、企業やコミュニティが完全に自由であることを意味するものではありません。

61. 法的ライセンスと社会的ライセンスの主な違いは、社会的ライセンスの「継続的」な性質であり、社会的ライセンスは（その付与方法を規定する明確なルールがないため）取得が難しく、撤回が容易かつ迅速であることです。

62. Emily Mullin, “Voice analysis tech could diagnose disease,” MIT Technology Review, January 19, 2017.

63. これは人間にとっても機械にとっても問題であり、スコットランドを舞台にしたテレビ番組をイングランドに向けて放送する際の字幕にも見ることができます。詳しくはJennifer Hale, “Scots in stitches after STV use subtitles to help viewers understand Glasgow accent in Ross Kemp: Behind Bars as he looks at life in HMP Barlinnie,” Scottish Sun, November 2, 2017.を参照ください。

64. 最近の話題の例としては、COVID-19の影響下における病院での人工呼吸器の割り当てという課題があります。詳細は、Robert D. Truog, Christine Mitchell, and George Q. Daley, “The toughest triage—allocating ventilators in a pandemic,” New England Journal of Medicine 382, no. 21 (2020): pp. 1973–75.を参照ください。

65. メルボルン出身の著者は、サンフランシスコに住んでいたとき、午後8時のレストランを電話で予約することができませんでした。"8 p.m. "という言葉をどのように表現しても、電話に出たレストランのスタッフは何を言っているのか理解できませんでした。そこで、19:30か21:00に予約を入れ、代わりに20:00を提示されたらそれを受け入れることにしたのです。

66. 医学研究は男性に偏っており、特に患者と医師の性別の不一致時に顕在化します。AIはこの偏りを増幅させ、例えば女性患者の心臓発作を誤診する可能性があります。一例として、Brad N. Greenwood, Seth Carnahan, and Laura Huang, “Patient–physician gender concordance and increased mortality among female heart attack patients,” Proceedings of the National Academy of Sciences 115, no. 34 (August 6, 2018): pp. 8569–74.を参照ください。

67. The problem of being seen to be “crying wolf.”

68. これは前述のトロッコ問題とリンクしていて、近接性の影響、モラルハザード、そして正しい選択はないという点を指摘しています。

69. ここで「差別」という言葉を使っているのは、システムが利益を提供するためには、人々を特定して異なる扱いをし、個人を差別する必要がある場合が多いことを受け入れなければならないからです。問題は、この差別が一部の個人にとって望ましくない結果をもたらす場合に発生します。

70. 大目に見る、同意する、承諾する、という気質。

71. 好意的に見ている、賛成している、喜んでいる。

72. Harmon Leon, “Whole Foods secretly upgrades tech to target and squash unionizing efforts,” Observer, April 24, 2020.

73. Alex Hern, “Anti-surveillance clothing aims to hide wearers from facial recognition,” Guardian, January 4, 2017.

74. 「働くことの質」とは、「良い仕事」を意味しています。それは、過度な強制やストレスがなく、労働者の生まれ持った特性や能力を最大限に活かし、仕事が労働者にモチベーション、新規性、多様性、自律性、ワーク・ライフ・バランスをもたらし、労働者が適切な報酬を得て、雇用契約を公正なものと考えている仕事のことです。重要なことは、良い仕事は労働者が自ら学ぶことを支援し、そうすることで3つのレベルで利益をもたらすということです。労働者は個人的な成長と仕事の満足感を得ることができ、組織はスタッフが解決すべき新しい問題や追求すべき機会を見つけることでイノベーションを起こし、地域社会全体は繁栄する組織と労働者を受け入れることで経済的な利益を得ることができます。このようにして、組織にとって生産的で持続可能な「良い仕事」は、働く人にとっても魅力的で充実したものとなります。また、より幸せな市民とより高い生活水準を手に入れるためには、良い仕事をより大きなコミュニティの価値観や規範と一致させる必要があるのです。詳しくは、Peter Evans-Greenwood, Alan Marshall, and Matthew Ambrose, Reconstructing jobs: Creating good jobs in the age of artificial intelligence, Deloitte Insights, July 18, 2018.を参照ください。

75. これは、企業、企業が運営するプラットフォーム、第三者、そしてプラットフォームに居住する消費者の間の曖昧な代理関係によって複雑になっています。多くの場合、これらの当事者間の境界は不明確です。

76. Sydney Bauer, “Trans travellers face ‘invasive’ airport security at Thanksgiving,” Thomson Reuters Foundation News, LGBT+, November 28, 2019.

77. オペレーショナル・コミュニティ・エンゲージメント

78. 戦略的なコミュニティ・エンゲージメント

79. 「技術的な詳細」とは、AIソリューションの技術的な仕組みと、倫理的な方法論やフレームワークの技術の両方からの詳細を意味します。

80. これは、TroposやI*などの手法で行われている要求モデリングを単純化したものと考えられますが、連結グラフの必要はありません。私たちは、実装に変換できるものではなく、ソリューションのアイデアを提供しているだけなので、切断された情報-判断-行動の連鎖のセットを持つことは十分可能です。

81. アルゴリズムは最新の形式主義に過ぎないのだから、この区別は本来ならば人間の意思決定と抽象的または形式化された意思決定の間にあるべきです。詳細は、James C. Scott, Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed (New Haven, CT: Yale University Press, 1999).を参照ください。

82. 書籍「Seeing Like a State」で、Scott氏は、世界を理解するためにアルゴリズム的なアプローチを採用することが、いかに世界をも変えてしまったかについて、優れた持続的な議論を展開しています。

83. Robodebtプログラムは請求書を送る判断をするのに、法律上、人間が必要だったと思われます。詳しくは、Dana McCauley and Rob Harris, "Lawyers warned federal government robodebt scheme was 'unlawful'," Sydney Morning Herald, February 6, 2020.を参照ください。

84. Ibid.

85. Clifford Geertz, “Chapter 1. Thick description: Toward an interpretive theory of culture,” in The Interpretation of Cultures, third edition (New York: Basic Books, 2017).

86. 人類学者は、フィールドワークにおいて、一緒に活動するコミュニティのメンバーを指して「インフォーマント」という言葉をよく使います。また、人類学者のコミュニティへの導入を仲介したり、コミュニティで人類学者を受け入れたりするゲートキーパーという考え方もあります。誰が人類学者を紹介し、受け入れ、仲良くするかによって、人類学者のコミュニティに対する見識が変わり、人類学者が誰と一緒にいるかによって、どのような情報が得られるかが変わります。そのため、コミュニティへの最初の導入とその仲介者は、その後の展開に大きな影響を与えるため、人類学者が多くの時間をかけて探し、計画するものです。時間が経つにつれ、コミュニティにはインフォーマントのネットワークができ、彼らが話す人々（インフォーマント）と、人類学者自身が（コミュニティのアクティブなメンバーとして）経験することの間で、コミュニティの要因を発見し、コミュニティを理解することができるのです。

87. Melanie (Lain) Dare, Jacki Schirmer, and Frank Vanclay “Community engagement and social licence to operate,” Impact Assessment and Project Appraisal 32, no. 3 (2014), pp. 188–97.

88. 著者は、これらのものを「コミュニティ・ファクター」と呼ぶことにしました。これは、論文の中で理解できる既存の用語を見つけるのが難しかったためです。研究対象となるコミュニティにとって意味のある名称を選ぶのが普通であり、典型的な選択肢としては、シナリオ、プリファレンス、ユースケース、ディメンション、あるいはユースケース・ディメンションなどがあります。残念ながら、この記事の文脈ではどれも使えませんでした。私たちがここで言いたいのは、私たちの言葉の選択は、学問としての人類学にとって典型的ではないということです。

89. アクターネットワークおよびアクターネットワーク理論（ANT）のより包括的な定義については、M. Callon, "Actor Network Theory," in Neil J. Smelser and Paul B. Baltes (eds.), International Encyclopedia of the Social & Behavioral Sciences (Oxford, UK: Pergamon, 2001), pp.62-66を参照ください。

90. 一般形態素解析とは、スイスの宇宙物理学者であり、カリフォルニア工科大学を拠点とする航空宇宙科学者であるFritz Zwicky氏が、多次元的で定量化できない問題群に含まれる関係性の総体を構造化して調査するために開発した手法です。この手法の概要については，Tom Ritchey, "General Morphological Analysis: A general method for non-quantified modeling," 2002.を参照ください。

91. “Matrixing,” in GMA terminology.

92. 制約充足とは、問題をモデル化または定義するために使用されるパラメータに課される条件を表す制約（ルール）のセットに対する解決策を見つけるAIおよびオペレーションリサーチの手法です。

93. Simon Wardley, “An introduction to Wardley (value chain) mapping,” Bits or pieces?, February 2, 2015.

94. トラッキング自体が問題である可能性もあり、結果的にトラッキングソリューションをいくら再構成しても、コミュニティに受け入れられることはないということを指摘しています。

95. これは、問題に対してあらかじめ定義された技術的解決策をコミュニティに持ち込むという一般的なアプローチとは正反対のものです。

96. Irfan Saif and Beena Ammanath, “‘Trustworthy AI’ is a framework to help manage unique risk,” MIT Technology Review, March 25, 2020.

97. Anna Jobin, Marcello Ienca, and Effy Vayena, “The global landscape of AI ethics guidelines,” Nature Machine Intelligence 1, no. 9 (2019): pp. 389–

あとがき

AIと共存する社会が実現しつつある今、私たちはAIとどのように向き合っていくのかという重要な岐路に立たされています。中でも「AIの倫理問題」は、AIに関する議論において重要な論点となっています。特に2020年12月の「GoogleがAI倫理学者を解雇した」というニュースは多くの関心を集めました。解雇されたのは、Googleの主な収入源である検索エンジンのサービスに潜む倫理的な問題を指摘したことが原因です。Google社内では、そんな彼女を擁護する動きが広がり、2,000人以上の従業員が彼女を支持する署名活動に参加しました。過去には2018年に3,000人を超える従業員が共同でAIの軍事利用に反対する署名活動を行った結果、Googleが国防総省にAIの提供を打ち切るまでに至ったこともあります。米大手IT企業に対して従業員が企業倫理を求めていく動きは、これからさらに広がっていくでしょう。

機械学習の過程では、膨大な入力データからAIが自らデータのパターンを認識していくのですが、人と人との外見や能力を区別する認識システムに機械学習を導入する場合、AIが結果的にジェンダーや人種を「差」として表現してしまうことがあります。AIを搭載したロボットやシステムが事故を引き起こした場合、その事故の責任を誰に帰属させるべきかという問題があります。相次ぐ自動運転車の死亡事故を受けて、医療用AIが医療ミスを起こしたときに責任をどうすればいいのかという議論が過熱しています。まだAIによる医療ミスが具体的に問題となった事例はありませんが、すでに医療現場においてAIの導入は進められています。議論が未熟なまま、医療用AIの実用化が進んでいく現状を危惧するような意見もあります。

このような中で企業だけでなく国家や国際機関も対応に乗り出しています。日本政府は18年にAIを使う際の7原則を策定し、「人間中心」を掲げました。欧州連合（EU）の欧州委員会も同様にAI倫理指針を公表しています。「機械にどこまで決定させるのか」「禁止すべき領域はあるか」といった観点もあり、科学技術の倫理を総合的に取り扱う国連教育科学文化機関（ユネスコ）が国際ルールづくりを進めています。しかし、自律機械に対する態度の違いは、社会的問題に対する世論の関心の大きさよりも、宗教的・文化的な背景が影響していることは想像に難くありません。その違いを無視して倫理を論じるべきではなく、今後、アナリティクスを推進する我々自身も考え続けなければいけない問題です。

翻訳監修者／執筆者プロフィール

毛利　研

マネジャー／
認定プロダクトオーナー（CSPO）

人工知能関連の実装能力、業務経験が豊富だけでなく、機械学習/深層学習に掛かるアルゴリズムの研究開発、同テクノロジーを活用したビジネスモデルの企画、戦略策定、アナリティクス組織立ち上げを強みとする。特に、自然言語処理およびマーケティングオートメーション領域に関して多くの経験を有し、アナリティクス組織への高度化支援やデータ分析活動の助言、データサイエンスの教育事業に従事。

大場　久永

マネジャー

金融機関を主軸として、アナリティクスに関わるアドバイザリー業務に従事。金融機関に対して信用リスク関連業務における機械学習及び深層学習の活用に関するプロジェクトに従事する他、AML領域におけるアナリティクス活用を推進している。直近ではサイバーセキュリティにおけるアナリティクス活用など幅広い領域での業務経験を有する。

執筆者プロフィール

Peter Evans-Greenwood | pevansgreenwood@deloitte.com.au

Peter Evans-Greenwood is a fellow at the Deloitte Australia Centre for the Edge, helping organizations embrace the digital revolution through understanding and applying what is happening on the edge of business and society. He has spent 20 years working at the intersection between business and technology. These days, he works as a consultant and strategic adviser on both the business and technology sides of the fence.

Rob Hanson | rob.hanson@csiro.au

Rob Hanson is a senior research consultant at the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia's national science agency. Hanson is a transdisciplinary researcher who works at the intersection of emerging technology and their policy implications. He has a professional background in technology, security, risk management, and strategic foresight. Hanson's current research focus includes data privacy, consumer data, and trust. As a PhD candidate, his thesis is titled "Fables for future technology."

Sophie Goodman

Sophie Goodman is an applied anthropologist who previously worked in customer-led strategy and service design projects as part of Deloitte Digital. She uses her experience and training in ethnographic research to help organizations better understand and act on the needs and experiences of their customers. Goodman has worked in research roles for an Australian university and a global workplace culture consulting firm as well as on customer and user experience projects.

Dennis Gentilin | degentilin@deloitte.com.au

Dennis Gentilin helps organizations build the infrastructure required to promote ethical behavior and drive sustainable performance. A unique career experience catalyzed his strong interest in ethics. His book, The Origin of Ethical Failures, won the textbook prize in the 2017 UK Chartered Management Institute Book of the Year awards. He is an adjunct fellow at Macquarie University and an honorary fellow at the Centre for Ethical Leadership.

Deloitte.

デロイト トーマツ

デロイト トーマツ グループは、日本におけるデロイト アジア パシフィック リミテッドおよびデロイトネットワークのメンバーであるデロイト トーマツ合同会社ならびにそのグループ法人（有限責任監査法人トーマツ、デロイト トーマツ コンサルティング合同会社、デロイト トーマツ ファイナンシャルアドバイザリー合同会社、デロイト トーマツ税理士法人、DT弁護士法人およびデロイト トーマツ コーポレート ソリューション合同会社を含む）の総称です。デロイト トーマツ グループは、日本で最大級のビジネスプロフェッショナルグループのひとつであり、各法人がそれぞれの適用法令に従い、監査・保証業務、リスクアドバイザリー、コンサルティング、ファイナンシャルアドバイザリー、税務、法務等を提供しています。また、国内約30都市以上に1万名を超える専門家を擁し、多国籍企業や主要な日本企業をクライアントとしています。詳細はデロイト トーマツ グループWebサイト（www.deloitte.com/jp）をご覧ください。

Deloitte（デロイト）とは、デロイト トウシュ トーマツ リミテッド（“DTTL”）、そのグローバルネットワーク組織を構成するメンバーファームおよびそれらの関係法人のひとつまたは複数を指します。DTTL（または“Deloitte Global”）ならびに各メンバーファームおよびそれらの関係法人はそれぞれ法的に独立した別個の組織体です。DTTLはクライアントへのサービス提供を行いません。詳細は www.deloitte.com/jp/about をご覧ください。
デロイト アジア パシフィック リミテッドはDTTLのメンバーファームであり、保証有限責任会社です。デロイト アジア パシフィック リミテッドのメンバーおよびそれらの関係法人は、それぞれ法的に独立した別個の組織体であり、アジア パシフィックにおける100を超える都市（オークランド、バンコク、北京、ハノイ、香港、ジャカルタ、クアラルンプール、マニラ、メルボルン、大阪、上海、シンガポール、シドニー、台北、東京を含む）にてサービスを提供しています。

Deloitte（デロイト）は、監査・保証業務、コンサルティング、ファイナンシャルアドバイザリー、リスクアドバイザリー、税務およびこれらに関連するプロフェッショナルサービスの分野で世界最大級の規模を有し、150を超える国・地域にわたるメンバーファームや関係法人のグローバルネットワーク（総称して“デロイトネットワーク”）を通じFortune Global 500®の8割の企業に対してサービスを提供しています。“Making an impact that matters”を自らの使命とするデロイトの約312,000名の専門家については、（www.deloitte.com）をご覧ください。

Member of
Deloitte Touche Tohmatsu Limited

bsi. ANAB ACCREDITED ISO/IEC 17021 MANAGEMENT SYSTEMS CERTIFICATION BODY
IS 669126 / ISO 27001