資料７．「治療においてＡＩをどのように開発、検証、実践するべきか」（AMA Journal of Ethics 誌・2018 年）

アブストラクト

人間の病理学者よりも成功率が高いと思われるＡＩは、がん細胞を発見する際に人間にとって代わるべきなのか、それとも人間を支援（augment）すべきであるのか。本論文では、ブラックボックス問題やオートメーションバイアス（医師が診断についてＡＩに頼り過ぎること）は患者の視点からすればそれほど問題ではなく、検査結果を適切に評価するのにＡＩについて熟知することが求められるということを論ずる。

ケース

A医師は地域の病院に勤めている病理学者である。ある日、彼女は朝からJ患者の生検標本を評価していた。J氏は53歳の女性で、乳がんの検査のためのセンチネルリンパ節生検をともなう腫瘍摘出手術を受けたばかりだった。担当の外科医は４つのリンパ節生検標本を作り、生検に回した。リンパ節生検結果はコンピュータに読み込まれ、数ギガバイトになる高解像度のイメージとなり、A医師が診断するためのJ氏のカルテに統合された。

A医師は異常な細胞を映した画像を探している最中、グーグル社が生み出したＡＩシステムを紹介する記事を読んだことを思い出していた。このシステムを使えば、ほんの数秒で、がんが疑われる細胞を画像から探し出すことができるらしい。その記事によれば、そのシステムによるがん細胞発見率は92.4%で、人間による発見率73.2%よりも優れている。しかもそのシステムは、機械学習によって発見率をさらに高めることもできる。

A医師は、自分の医療チームがJ氏のような患者を助けるためにスクリーニングツールとしてこのシステムを活用できればよいと思った。A医師の同僚たちは、この技術によって患者の生検標本から異常を発見する確率は高くなるだろうという点では意見が一致していたが、そのシステムから偽陽性や偽陰性が出てきてくる可能性も否定できないのでそれに投資するよう病院に助言することを躊躇する者も何人かいた。また、何人かはオートメーションバイアスも問題だと考えていた。この問題は、利用者がＡＩプログラムを活用することで簡易化された仕事の流れに慣れてしまい、その結果、そうした臨床の意思決定支援システムに過度に依存してしまう、というものである。医師Aのチームは、これらの問題について次回の病理学分野のミーティングで協議することになっ

ている。

コメンタリー

人間によるがん細胞発見率に比べてＡＩの方が優れているという点で言えば、ＡＩは人間が病理学的にがん細胞を発見することを支援するものとして活用できる、という主張を裏づけるように思われる。さらには、人間の肉眼よりも優れているから、ＡＩプログラムを使用することで病理学者のトレーニングに役立つとさえ主張できるかもしれない。人間では看過する恐れのあるがん細胞を発見できるシステムを使わない手はないだろう。これがその活用を促すひとつの理由である。こうしたシステムが存在することを知っていても導入せず、当該の病院の病理学者ががん細胞を見落とし、その結果患者が死んでしまうような場合に、その病院が訴えられるというケースは想像に難くない。その一方で、このシステムには検討すべき倫理的問題が２つある。

ひとつはブラックボックス問題である。ＡＩプログラムはニューラルネットワークアルゴリズムに基づいているので、それがどのような仕方でがん細胞を発見しているのか仕組みが分からない。しかしながら、もしそのプログラムが病理学者に取って代わるのではなく病理学者の支援として活用されるのであれば、ＡＩはがん化した細胞がどれなのかに関する知識を得るうえで病理学者の助けになるのではないか。病理学者の仕事には、がん細胞（あるいは通常の細胞）が共通してもつ視覚的特徴を同定したり、ＡＩプログラムによってがん細胞だと同定される傾向にあるがん細胞の特徴を視覚化したりすることが含まれうるだろう。こうした仕事を実行することで、ブラックボックス問題は軽減されるだろうし、達成される透明性からして病理学者や患者がＡＩシステムにより安心して依拠できるように仕向けることもできるかもしれない。

他方で、ＡＩプログラムが広く活用され、がん細胞発見率が現在の 92%からほぼ 100%にまで向上したと想定してみよう。ＡＩがどのようにして発見しているのかについて、私たちはそれほど気にするだろうか。もちろん医療従事者の中には関心を抱く人もいるだろうが、がん細胞の有無によって人生を左右される患者にとっては果たして問題なのだろうか。私たち筆者は問題ではないだろうと考える。なぜなら、このＡＩプログラムが実行する課題は、他のＡＩプログラムと対照的に、事実として黒か白かの判断を下すことだからである。つまり、細胞にがん細胞が事実として含まれていれば、ＡＩプログラムはそれを発見する、ということである。私たちが気にするのは、それがどれだけ正確にがん細胞を発見できるのかである。もしそれが 100%の確率で発見でき、人間ではそれに到底追いつけない状態であれば、

ＡＩが病理学者に取って代わることもありうるだろう。実際のところ、ＡＩは仕事はきわめて早いうえ、疲労などによる人的ミスから誤った判断を下すということもほぼない。

ここで主張しているのは、仮にＡＩプログラムによってがん細胞が発見される確率が 100%であるならば、ブラックボックス問題は倫理的にも臨床的にも重要ではなくなる、ということである。しかしその一方で、ＡＩプログラムが物議を醸すと思われるような仕方で人間の人生に影響を及ぼす可能性がある場合には、ブラックボックス問題はそれでも重要であり続けるだろう。たとえば、ＡＩによる自動車の自動運転がそれである。そこでのブラックボックス問題は、自動運転の車がなぜ運転に関する判断を下しているのかを理解することに関係するのであって、それが倫理的に許容できない理由を知ることでない。すなわち、私たちの大半は、ＡＩ技術や機器を安心して使用する前に、それが下す「判断」に関する基礎的な知識を得たいのである。

もうひとつの問題はオートメーションバイアスである。一般的に言えば、これは、医療従事者がかつて担っていた仕事がＡＩプログラムの仕事となる場合に生じうるある種の満足感を指す。もしＡＩプログラムがほぼ 100%の成功率をキープするのであれば、オートメーション自体に倫理的・臨床的問題があるとは思えない。しかし、その成功率がそれより低い場合―たとえば上述のケースのように 92%である場合、当該のプログラムには質のよいインプットがあるということに、私たちが確信もっていることが重要である。上述のケースで言えば、おそらくそれは、ＡＩプログラムが多くの年齢層と人種からなる女性患者を横断的に調べて「学習した」ということを意味しているだろう。多様なインプットが確保されれば、倫理的にまた臨床的に最も重要となるのは、ＡＩが人間の病理学者よりもがん細胞発見率が高いという点である。

ＡＩプログラムも人間も 100%の成功率が達成できないとして、その場合の目標ができる限り最も正確な診断を下すことだとすれば、人間の病理学者の知識と技能がＡＩによって支援される（augmented）ときに最も高い成功率が出る可能性はある。おそらく、何かしらミスが生じた場合には、ＡＩプログラムは人間によるミスよりも異なる類のミスを出すだろう。とすれば、病理学者とＡＩ両方の方法を用いることが最も正確な診断を出すように思われる。

架空か実際か

ここで扱っているケースが架空のシナリオとして理解するならば、提示される事

実は額面通りとして捉えることができる。実際、先のコメンタリーではこの理解に基づいて書いている。他方で、それを実際のシナリオとして理解するならば、何らかの分析が実施される前に、提示される諸事実が実際に妥当であるのかを検証するためにより慎重なプロセスがなければならない。言わずもがな、ＡＩに詳しくない人にとってはＡＩの技術的な詳細はとりわけ問題となる。たとえば、A 医師は、ＡＩアルゴリズムの中身に関する事実を考慮する前に、グーグル・ブレイン・プロジェクトに関するいくつかの事実を調べるべきである。

第一に、現時点では、グーグル・ブレイン・プロジェクトは、がん細胞の発見のためのアルゴリズムとして使用されることを念頭に置いていない。グーグルＡＩプログラムによれば、「トレーニングモデルは、興味深い研究と実際のプロダクトとの橋渡しをするための数あるステップの第一段階に過ぎません。臨床的な妥当性から規制当局による承認に至るまで、「基礎から応用」までにはかなりの距離があります。とはいえ、私たちは見込みあるスタートを切ったところです。私たちは私たちが作り出したプログラムを共有することで、この発展に拍車をかけることができるでしょう」。開発者側は自分のプロダクトをできる限り見栄えがよいように示しがちであることを前提にすれば、自分のプロダクトに対する開発者側の評価は私たちの期待値の上限として提示されていると捉えるのが賢明であろう。

第二に、人間による 73.2%の発見率は、実際のところ特定の画像を見た１人の病理学者が出した結果であるので、その数値が人間一般の限界を示すわけではないということは明記しておかねばならない。こうした限界を定めることは、技術の開発者の重要な課題となるだろう。人間の肉眼による発見率が高くなるのであれば、病理学者の仕事をＡＩで置き換えることの説得力はなくなってしまう。

第三に、ＡＩプログラムが時間とともに改善していくことは可能である一方で、現時点では従来言われているようには改善が見られない。Y. Liu 氏によれば、「結果の安定性と再現可能性が重要視されるような医療においては、むしろ「固定された（frozen）」トレーニングモデルの方が好ましい」。すなわち、現状の信頼度（「固定された」ＡＩプログラム）には、問題となるケースに関する事実を考慮し検討する際に必要な臨床的・倫理的価値がある、ということである。しかしこれは他方で、ＡＩプログラムが今後学習していかないのであれば、機械学習のプロセスに倫理的価値を付与すべきではない、ということも含意する。

最後に、本論文でのケースでは数値は明確に示していないが、ＡＩプログラムを用いたがん発見においても偽陽性や偽陰性が実際のところ報告されている。ＡＩプ

ログラムによって偽陽性が出る確率は１画像当たり 0.0001%以下であり、きわめて小さい。偽陰性の確率は本論文でのケースにおける 92.4%の精度に示唆されているように、ＡＩシステムによるそれは 7.6%である。こうしたデータが存在するのであれば、A 医師はこれらの確率を考慮に入れてＡＩプログラムを活用するかどうか考えることができる。

倫理的観点からすると、上述のケースにおけるＡＩプログラムは十全なバリデーションのプロセスを経ていないのではないのか、という懸念が残る。たとえば、そのＡＩプログラムの成功率は、たった１枚の画像を用いたたった１人の病理学者の成功率と比較されているだけである。この比較だけでは、当該のＡＩプログラムの成功率がある特定の病院の診断医たちのそれよりも優れているという、確たる証拠にはならない。その一方で、十分な質のデータを用いてＡＩが叩き出す 92.4%の成功率はそれでもとても高い水準であり、疲労や時間の制限のある人間の病理学者がなかなか超えることのできない水準である。この点を加味すると、ＡＩプログラムの活用はひとつの賢明な手段であるように思われる。

本論文でのケースに関する事実やＡＩプログラム研究について誤解が生じる恐れがあることも考慮して、ここでは、倫理審査委員会には医療や倫理の専門家が含まれなければならないのと同じように、医療におけるＡＩを活用する場合の倫理審査にはＡＩの専門家も含まれる方がよいと提案したい。

（仮訳：山本圭一郎）

著者：Anderson M, Anderson SL.
原題：How Should AI Be Developed, Validated, and Implemented in Patient Care?
出典：AMA J Ethics. 2019 Feb 1;21(2):E125-130. doi: 10.1001/amajethics.2019.125.