特集 **使えるAI**

シンセティック・メディア
AIによるメディア制作の新潮流

長谷佳明

CONTENTS

要　約

1　企業は商品やサービスの説明手段として、動画に高い関心を持っている。また、2020年は新型コロナウイルスの蔓延により、動画の需要はより一層高まった。今後、従来のスタジオ収録に代わる生産性の高い動画制作技術として期待されるのが「シンセティック・メディア」である。

2　シンセティック・メディアとは、リアルな音声付き動画をAIによって作り出す動画合成技術である。事前に用意したテキストと人（映像、音声）のデータからAIが学習し、その人が本当に話しているような動画の生成が可能である。

3　新華通訊社は2018年からシンセティック・メディアによる「AIアナウンサー」を活用し始めている。AIアナウンサーは多言語化が容易であり、ミスがないため撮り直しが不要で、24時間対応も可能など、ニュース制作にとって画期的な技術といえる。

4　シンセティック・メディアによって、将来的に動画制作の一部は、AIによる生成に置き換わる。人のデジタルツインが使われるようになり、動画との対話が可能となる。

5　シンセティック・メディアを悪用したものが、ディープフェイク（偽動画）である。技術保有企業や研究機関は、関連するソフトウエア技術の流出を防止し、本人の明示的な同意なしにシンセティック・メディアによる再現を禁止するなど、独自に対策を始めている。日本国内でも、悪用防止に向けた倫理ガイドラインの策定が急務である。

Ⅰ シンセティック・メディアとは何か

1 近年の動画共有サービスの変遷と新型コロナウイルスの影響

2000年代半ばのYouTubeの登場に始まり、10年代のNetflixなどの動画配信サービスの普及、さらには近年のTikTokのようなショートビデオの人気を経て、動画はわれわれの生活にすっかり溶け込んでいる。この結果、企業は従来の紙の説明書に加え、説明用の動画も用意するようになっている。

スマートフォンと高速データ通信サービスの大衆化により、移動中や待ち時間などの隙間時間に動画の視聴が可能となり、新たなビジネスも開花している。たとえば、アリババグループの「タオバオライブ」に代表される「ライブコマース」（ECと動画を融合したサービス）である。

ライブコマースはテレビさながらの演出による「エンターテインメント性」に加え、配信者と視聴者がリアルタイムでコミュニケーションする「双方向性」、オンライン決済サービスの「利便性」を融合した新たなサービスである。日本でも、大手百貨店を中心に取り組み始める企業が登場するなど、動画はビジネスでも高い関心を集めている。

20年から世界を襲った新型コロナウイルスの蔓延により、動画への新たな需要も喚起されている。たとえば、オンライン教育である。海外の先進的な大学などで進められてきたオンライン教育が、コロナ禍により対面授業が難しくなったあらゆる学校で求められるようになった。Zoomのようなビデオ会議システムを使い、ライブ配信によって対応した学校も少なくない。中には録画した授業を再利用することもあるが、コンテンツとしての完成度は必ずしも高くはない。言い間違いやスライドの操作ミス、マイクの思わぬミュートなど、慣れない作業に戸惑う教員が少なくないためである。このため視聴に耐え得るコンテンツを作り出すには、再度の収録や編集が必要になり、需要の増加に応えられる状況ではない。

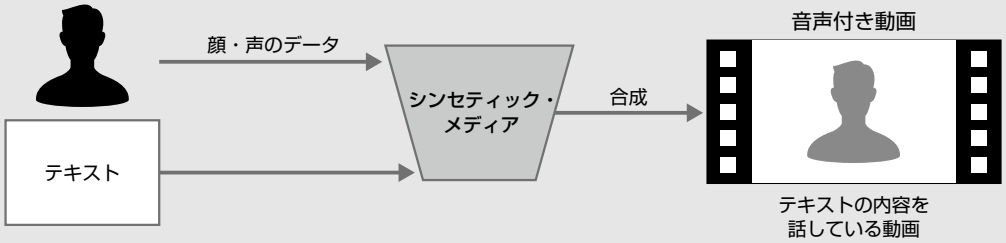
同じく非対面での対応を余儀なくされているのが小売業である。コロナ禍以前に見られたスーパーでの営業員による商品紹介や実演販売はすっかり姿を消し、代わりに見かけるのがデジタルサイネージの活用である。イオングループ傘下のユナイテッド・スーパーマーケット・ホールディングスでは、デジタルサイネージとカメラを組み合わせ、年齢層や性別に合わせた商品動画を配信して接客する取り組みを20年３月から始めている。このようにデジタルサイネージと動画を組み合わせた取り組みは増加するだろう。

今後、新商品の販促のためのコンテンツ制作や、値引き販売などに応じたタイムリーな動画制作への要望も高まるため、安価に、より早く動画を作り出せる技術へのニーズは、ますます高まると予想される。

2 シンセティック・メディアとは何か

動画をめぐる新たな技術開発も始まっている。それが「シンセティック・メディア」である。シンセティック・メディアとは、実際にカメラを使って撮影したかのようなリアルな音声付き動画をAIによって作り出す動画合成技術である。事前に用意したテキストと

図1　シンセティック・メディアによる音声付き動画の生成



人（映像、音声）のデータからAIが学習し、その人が本当に話しているような動画の生成が可能である（図１）。

中国の新華通訊社は、2018年、テキスト原稿を与えると、男性アナウンサーが本当にその原稿を読み上げているような、リアリティの高い動画を自動生成するシステムをシンセティック・メディアによって開発し、24時間いつでもニュース制作できる体制を整えた。技術開発には、中国の検索エンジン大手Sogou（搜狗）が協力した。Sogouは、「リップリーディング」と呼ばれる人の口元の動きから発話内容を予測する画像認識技術を以前から研究しており、それを応用し、合成動画の口元を再現する技術をいち早く開発した。

Sogouは19年、女性AIアナウンサー「Yanny」を披露している（図２）。Yannyはカンファレンス会場でSogouの決算発表を行うなど、活躍をリアルの場にまで広げている。本物のアナウンサーのように状況に応じた臨機応変な対応は難しいが、シンセティック・メディアの作り出すAIアナウンサーは、言い淀むことがなく、多言語化も容易である。

図2　AIアナウンサー「Yanny」

出所）https://www.facebook.com/103400854571373/videos/2858771370897031/

Ⅱ　技術詳細

1　従来型のCG制作とシンセティック・メディアとの違い

合成動画を作り出す技術には、映画製作で使われてきたCG技術もある。たとえば、米国のデジタル・ドメインは、2013年に台湾の歌手テレサ・テンの没後20周年を記念し、フルCGによる合成動画を制作した。多数のCGクリエイターが５カ月もの期間をかけ、約16億円の巨費を投じたハンドメイドCGの集大成であった。

一方、シンセティック・メディアはAIを使った合成動画システムであり、質や表現力こそCGには及ばないものの、データさえあれば容易にリアリティの高い動画を生成でき

る。たとえば、イスラエルのスタートアップであるCanny社は、19年にインターネット上で公開されていたわずか21秒の会話データを使い、フェイスブックのCEOであるマーク・ザッカーバーグ氏を使った架空のニュースを制作している。シンセティック・メディアを活用して制作されたこの動画は非常にリアリティが高く、一見すると本物と見分けがつかないほど精巧である。

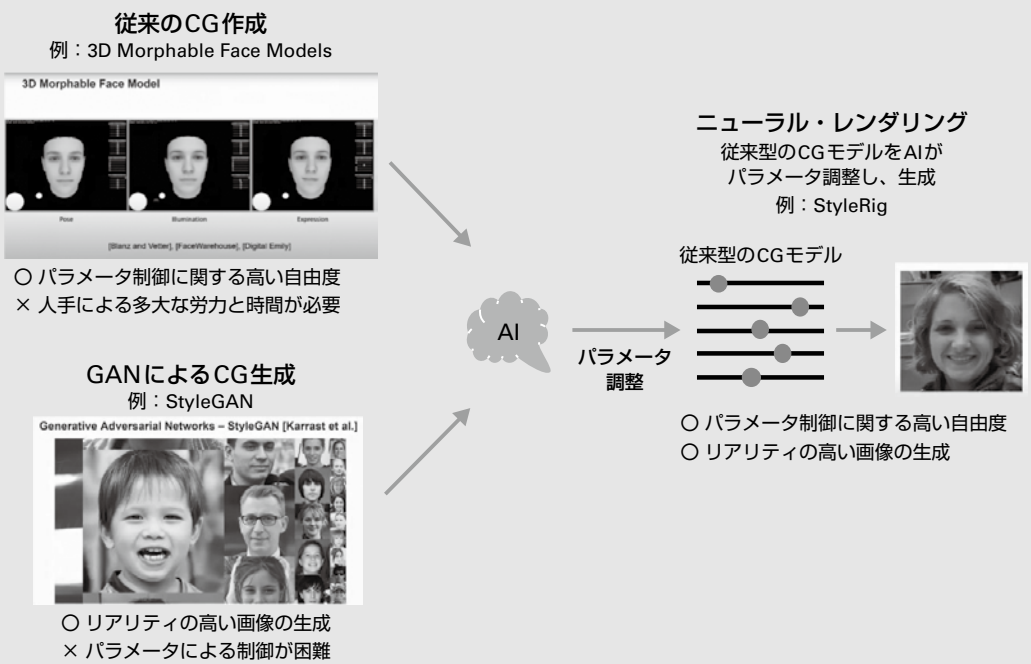
シンセティック・メディアは、合成動画の制作プロセスを「職人によるハンドメイド」から「AIによるオートメーション」に換えようとしている。

2　ニューラル・レンダリングの発展

シンセティック・メディアの技術の基礎になっているのが、「ニューラル・レンダリング」である。ニューラル・レンダリングとは、人間が精巧なCGモデルを設計して画像を生成する代わりに、AIが対象物の三次元構造を予測して画像を生成する技術である。図3に、「StyleRig」と呼ばれるニューラル・レンダリングを使った手法を示した。

従来型のCG技術の中でも、人の顔の表現に強みを持つ技術が「3D Morphable Face Models」である。このモデルは人の顔の構造を忠実に再現したもので、顔の凹凸を微妙に調整し、肌の上に色を重ねることでリアリティの高いCGを作り出せる。しかし、人手による作業が多いため、多大な労力がかかるのが欠点である。一方、「GAN（Generative Adversarial Network）」と呼ばれるAIモデルを応用した技術もあるが、この技術だけでは生成される顔の特徴や表情を細かく制御す

図3　ニューラル・レンダリングを使った顔生成技術「StyleRig」



出所）CVPR2020 Tutorial, Neural renderingを参考に作成

るのが難しい。

シンセティック・メディアでは、従来型のCG技術とディープラーニングを組み合わせたニューラル・レンダリングと呼ばれるAI技術が使われている。ニューラル・レンダリングでは、従来型のCG技術で使われる顔の構造モデルを使いながらも、その調整作業をディープラーニングで置き換えている。この結果、あたかも本物の人間が話していると見間違うほどの写実性の高い動画を容易に作り出すことが可能となったのである。

3　シンセティック・メディアの実現レベル

シンセティック・メディアの制作には最新のAI技術が使われており、現在もなお改善が続けられている。まばたきや音声に合わせて口を動かしたり、特定の人の声に似せたりなど、既に顔や声の再現性は高く、本物の人間とほとんど区別がつかない。

一方で、体の動きの再現については研究途上にあり、特定のパターンを演じることはできるものの、個人の癖を違和感なく再現するのは難しい。

Ⅲ 活用事例

1　Samsung NEXT Ventures「デジタルツイン」によるプレゼンテーション

サムスン電子でビジネスのインキュベーションなどを担うサムスンネクストは、2020年8月、シンセティック・メディアの説明動画「Synthetic Media Landscape 2020」を公開した。この動画自体がシンセティック・メディアを使って制作されており、作成者のIskender Dirik氏にそっくりな「デジタルツイン」が本人に代わってプレゼンテーションを行っている（図4）。

これはあらかじめ話す内容をテキストで準備し、シンセティック・メディアによって本人そっくりのデジタルツインが講演スライドに合わせてテキストを読み上げるというものだ。人が実際に話す場合と比べ、言い間違いなどによる撮り直しがなく、目線も自然で説得力のある印象を受ける。シンセティック・メディアによって人のデジタルツインを容易に制作できるようになれば、オンライン教育向けのコンテンツ制作などにも生かすことができるだろう。

2　WPP　コンテンツの多言語化、パーソナライズ

世界的な広告代理店グループであるWPPは2020年、社内教育用コンテンツの制作にシンセティック・メディアを活用した。同社は全世界に多くの関連企業を抱えている。そのため、たとえばコンプライアンス研修一つをとっても、役割に応じたシナリオを準備し、多数の言語に対応しなければならず、制作費が数千万円に及ぶこともあった。

同社はシンセティック・メディアの持つテキストからリアリティの高い会話動画を生成する技術を応用し、費用を1000万円に抑えることに成功した。制作に協力したのは英国のスタートアップSynthesia社である。制作されたコンテンツは、パーソナライズされており、一人一人の名前で呼びかけ、社員に合わせたシナリオで進められるため、学習効果の向上も期待できる。

WPPは、シンセティック・メディアの評価を重ねることで、将来的には自社の広告ビジネスやメディア事業での活用を検討していくものと推測される。

3 映画『Welcome to Chechnya』デジタルベール

2020年に公開されたデイヴィッド・フランス監督によるドキュメンタリー映画『Welcome to Chechnya』では、顔をモザイクで隠す代わりに、シンセティック・メディアによって別人の顔に置き換えている。この映画は、ロシア連邦に属するチェチェン共和国内のLGBTに対する厳しい迫害にスポットを当てた作品である。

これまで使われてきたモザイク技術では出演者の表情が分からず、ドキュメンタリーとしての情報量が落ちてしまう。一方で、モザイクがない場合、映画公開により出演者の身元が公のものとなり、迫害組織の脅威にさらされることになり得る。そこで本映画では、シンセティック・メディアを活用し、出演者の顔をニューヨークなどに住む協力者の顔に変える「デジタルベール」を使った。この結果、出演者の顔からにじみ出る心情を正確に残すことができた。

4 ソニー「イマーシブリアリティ・コンサート」VRエンターテインメント

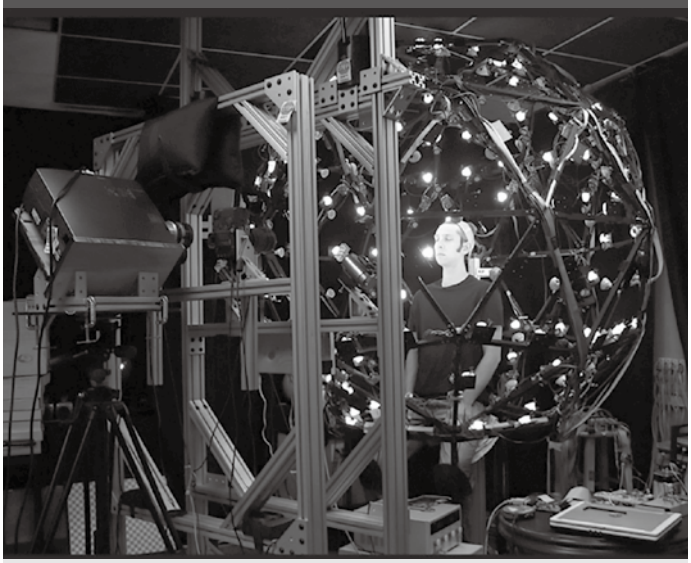
特別なデータを用意すれば、映像作品にも堪え得る極めて高精細な合成動画の制作も可能となる。たとえば、ソニーが2021年のCESで発表した米国の人気歌手マディソン・ビアーによる「イマーシブリアリティ・コンサート」である。

音声にはマディソン・ビアーの生歌が使われているが、その姿はシンセティック・メディアによって生成されている。シネマクオリティにまで映像品質を高めるため、人の顔や身体に関する大量の画像データの収取には南カリフォルニア大学が開発した「ライトステ

図4 Synthetic Media Landscape 2020

出所）https://www.syntheticmedialandscape.com/

図5 ライトステージ



出所）https://vgl.ict.usc.edu/LightStages/images/Avatar/Final/IMG_5790.JPG

ージ」と呼ばれる装置が使われた（図５）。ライトステージは、映画『アバター』など数々の映画にも使用された装置で、球体状の装置の中に被験者を入れ、周辺に張りめぐらしたLEDライトを微妙に調整しながら、顔の陰影による変化をカメラでとらえてデータ化する。

ライトステージによって取得したマディソンの精細な顔データと本人の動きをモーションキャプチャで数値化し、AIが「デジタルツイン」のマディソンをコンピュータ上に再現する。

このコンサートは、「イマーシブ＝没入型」というタイトルが示す通り、VRゴーグルを装着して間近で視聴できるだけでなく、通常のコンサートでは体験できない仕掛けが施されている。たとえば、視聴者と歌手の目線が一瞬合う体験である。実際のコンサート会場でも、ファンであれば一度は願う瞬間を、シンセティック・メディアによって3D合成した仮想空間上で再現している。

現在のシンセティック・メディアの技術では、シネマクオリティにまで品質を向上させるためにはAIに与えるデータに特別な対応が必要になるが、エンターテインメント分野での活用も有望といえるだろう。

Ⅳ 今後の展望：シンセティック・メディアの活用に向けて

シンセティック・メディアによって将来的に動画にかかわる三つの変化が起こると予想される。

1 動画の制作方法の変化

一部の動画はAIによって生成されるようになり、格段に人手がかからないものとなる。既に新華通訊社が開始しているように、ニュースとシンセティック・メディアは相性が良い。シンセティック・メディアは原稿さえあれば、一字一句間違えることのない音声付き動画を、24時間いつでも作り出せる。報道や施設案内、商品説明など、特別な演出や演技よりも正確性が求められるシーンでは、動画制作の担い手はAIに置き換わるだろう。

シンセティック・メディアには、音声付き動画の活用の裾野を広げる効果も期待できる。たとえば、不動産賃貸業である。既にインターネット上で物件の間取りや写真を確認できるようになってはいるが、テキストと写真だけのサイトはPCで閲覧する場合は問題ないものの、スマートフォンでは使い勝手が悪い。

シンセティック・メディアを使えば、物件一つ一つに合わせた説明用の動画を容易に作り出せる。シンセティック・メディアは、費用対効果の面で動画制作を諦めていた企業が動画を利用するきっかけとなるだろう。

2 キャストの変化

動画は、人が演じるだけでなく、用途によって人とデジタルツインを使い分けるようになる。実在する人の姿や声からシンセティック・メディアによって作り出されたデジタルツインが、本人の代役となる可能性がある。たとえば、著名人のデジタルツインが自分の誕生日を祝ってくれるようなメッセージ動画を作り出すサービスが登場するかもしれない。

エンターテインメント向け動画は、将来的に本人の代役ではなく、次第に「クリエーション（創造）」といえる段階に進む可能性もある。たとえば、自分の望む出演者を使ったオリジナルドラマの制作である。権利関係の調整が済めば、この世を既に去った俳優と現在の有名俳優を使った作品など、個人の望む「夢の共演」も実現するだろう。

著名人は、実際にテレビ番組などに出演する傍ら、自身のデジタルツインの権利を維持管理し、シンセティック・メディアによって作り出されたコンテンツからも出演料を得られる可能性がある。シンセティック・メディアは、キャストにとってデジタルツインという新たな「出演機会」を生み出す。

3　役割の変化

シンセティック・メディアによって音声付き動画がほぼリアルタイムで生成可能となれば、動画は視聴するだけのものから会話可能なメディアへと変化する。

「Amazon Echo」や「Google Home」などのスマートスピーカーが家庭で使われるようになった。音声対話により操作し、音楽を楽しんだり、ニュースや天気予報を確認したり、ラジオも聴けたりする。ディスプレイ付きの製品もあるが、アマゾンのアレクサをはじめ、スマートスピーカーのエージェントは声だけを持ち、容姿はない。しかし、シンセティック・メディアを使えば、好みのタレントの容姿と声を持つエージェントがディスプレイ越しにユーザーに話しかけることも可能となるだろう。「表情を持つアレクサ」の誕生である。2020年、これを予感させる出来事が米国で起こった。アメリカンフットボールの優勝決定戦「スーパーボウル2020」の中で放映されたアマゾンのCMである。アレクサが精悍な男性の姿となって女性の前に現れ、ユーザーの生活に寄り添うというイメージCMで、まさに「表情を持つアレクサ」の登場を連想させるものであった。

近年は、タブレット型のカメラを備えたスマートスピーカーもある。このようなタイプのスマートスピーカーであれば、カメラに映り込む人間の表情をとらえることも可能だ。もし、話しかけた相手が喜びの表情をしていれば、シンセティック・メディアによって作られたエージェントは、同じく笑顔で話しかけて共感を示すなど、これまでにない体験も実現する。将来的に、動画はAIなどで作られた知的なシステムのインタフェースへと昇華していくだろう。

V　課題

1　悪用への懸念：ディープフェイク

著名人の動画を加工したり、まったく別の動画を著名人の顔に置き換えたりした偽動画「ディープフェイク」が問題となっている。これはシンセティック・メディアを悪用したものであり、かつてはパロディの域を出なかったが、技術の進化に伴い、本物と区別がつかなくなりつつある。

日本でも2020年、アダルト動画の顔を著名なアイドルの顔に置き換えるなどしたコンテンツをインターネットに公開した男性が、名誉棄損と著作権法違反で逮捕されている。

シンセティック・メディアの技術が犯罪組織の手に渡れば、政治家に似せた偽動画をSNS上で拡散し、金融市場をコントロールし

たり、民衆を扇動したりしてデモや暴動を引き起こす可能性もある。シンセティック・メディアは、将来に向けた技術発展と同時に、そのリスクを早期に見極め、悪用を防ぐための仕組みを整える必要がある。

2　関連技術の流出防止や倫理ガイドラインの必要性

一般にAIの研究論文では、アルゴリズムを実装したソースコードや学習に使われたデータセットを合わせて公開し、再現性を証明する。しかし、シンセティック・メディアにかかわるAI技術は、研究者の間で早い段階からディープフェイクへの悪用が懸念されていたため、関連するソフトウエアのソースコードを非公開とするなどの対策が講じられている。

また、Synthesia社やサムスンネクストなど、シンセティック・メディアのビジネス活用を進める企業の中には、独自に倫理ガイドラインを設けている企業もある。たとえば、有名・無名を問わず、本人の許諾なく、いかなる公開情報もシンセティック・メディアに使わないこと、生成される動画が社会的、経済的に混乱を招くようなフェイクニュースに使用されないようにすることなどである。Synthesia社は、ソフトウエアを「販売」するのではなく、SaaS（Software as a Service）として提供し、生成されるコンテンツに問題がないかチェックしている。

ディープフェイクに関連する法規制の制定も始まっている。中国では、2019年頃にマッチングサービスを手掛ける陌陌科技（Momo Technology）が開発した「ZAO」と呼ばれるディープフェイクアプリが流行した。個人の娯楽の範囲で動画の顔を他人にすり替えて楽しむアプリであったが、急速にSNS上で広がり、著作権侵害にあたるものや著名人の顔を使った悪質な投稿が目立つようになった。

中国のサイバーセキュリティやインターネットを監視する政府機関である中国国家インターネット情報弁公室（Cyberspace Administration of China、CAC）は、19年11月、「ネットワーク音声・動画情報サービス管理規定[注]」を制定し、シンセティック・メディアのようなAI技術によって合成された動画に関するルールを明記した。他国に先駆けた「ディープフェイク対策法」といえるもので、20年1月から施行されている。

19からなる条文のうち、第11条ではディープフェイクを念頭に置いており、AIなどの新技術を使用して虚偽のニュースを作成し、インターネット上で公開することを禁じている。また、AIによって合成された動画に対しては、「AIによって生成した旨」を明記することを義務づけている。

また、第13条では、インターネットサービスプロバイダーに対して、AIなどによって作られた虚偽のニュースを発見した際には、関連する情報を報告するよう求めている。このほか、本規定では、「深層学習」というAIを意味する用語を度々引用し、ユーザーとプロバイダーにディープフェイクなどへの適切な対処を求める内容になっている。AIの社会実装で先行する中国ならではの素早い対応といえるだろう。

シンセティック・メディアは将来有望な技術であり、動画の制作方法、キャスト、役割をも大きく変える可能性を持っている。一方で、悪用へのリスクが伴うのも事実である。

今後、日本でも、シンセティック・メディアをビジネスに生かす取り組みやシンセティック・メディアを創り出すサービスが登場するだろう。政府には新たな音声付き動画生成技術としての産業振興と、ディープフェイクに代表されるリスクへの適切な対処という、推進と規制のバランスが求められる。一方、シンセティック・メディアを取り扱う企業は高い倫理観を持ち、生成されるコンテンツに対する責任が伴う点を忘れてはならない。

注

http://www.cac.gov.cn/2019-11/29/c_1576561820967678.htm

著者

長谷佳明（ながやよしあき）
野村総合研究所（NRI）IT基盤技術戦略室上級研究員
専門は人工知能、ロボティクス、IT基盤技術、開発技術／開発方法論など