The Ethics Of AI: How To Avoid Harmful Bias And Discrimination

Build Machine Learning Models That Are Fundamentally Sound, Assessable, Inclusive, And Reversible

by Brandon Purcell

February 27, 2018

Why Read This Report

While the potential existential threat of artificial intelligence is well publicized, the threat of biased machine learning models is much more immediate. Customer insights (CI) pros must learn how to identify and prevent harmful discrimination in their models, or businesses will suffer reputational, regulatory, and revenue consequences.

Key Takeaways

Biased Models Are Bad For Businesses
Most harmful discrimination is unintentional, but that won't stop regulators from imposing fines or values-based consumers from taking their business elsewhere.

Data Determines Discrimination
CI pros must defend against algorithmic or human bias seeping into their models while cultivating the helpful bias these models identify to differentiate between customers.

Ensure Your Models Are FAIR
CI pros should strive to create models that are fundamentally sound, assessable, inclusive, and reversible to protect against harmful bias.

The Ethics Of AI: How To Avoid Harmful Bias And Discrimination

Build Machine Learning Models That Are Fundamentally Sound, Assessable, Inclusive, And Reversible

by Brandon Purcell
with Srividya Sridharan, Fatemeh Khatibloo, TJ Keitt, Jennifer Wise, Christopher McClean, and Christian Austin

February 27, 2018

Table Of Contents

Related Research Documents

Brand Resilience: Understanding Risk Managers' Key Role In Protecting Company Reputation

Customer Empathy: Three Essential Strategies For Cultivating It In Your Organization

The Values-Based Consumer

FORRESTER®

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

Biased Models Harm Your Customers, Your Brand, And Your Business

The road to hell is paved with good intentions. Just ask Google. The company was a pioneer in image analytics, training a deep neural network to automatically classify users' images in Google Photo. Its model was quite impressive, recognizing a variety of objects like cars and skyscrapers. Then in June 2015, a black Haitian-American programmer in Brooklyn uploaded images of himself and a friend to Google Photo and the model labeled them as "gorillas."[1] Google had not intended to create a racist model, but that doesn't matter — unintentional racism is still racism. And it's not just unintentional racism that can seep into models. By their very nature, machine learning algorithms can learn to discriminate based on gender, age, sexual orientation, or any other perceived differences between groups of people. CI pros leading data science teams who unintentionally create biased models risk:

- **Reputational damage to their brand equity.** Companies that develop and deploy models that make harmful discriminatory decisions will inevitably face public backlash, leading to erosion of brand equity.[2] Amazon learned this when it rolled out Amazon Prime same-day delivery to certain ZIP codes in 27 US cities in 2015. A Bloomberg report found that in many of these cities, such as New York and Chicago, predominantly black ZIP codes were excluded from the service.[3] Amazon claimed that the decision was not based on demographics, but the headlines told a different story and the brand suffered from a torrent of negative press.

- **Regulatory fines and legal action.** Models with harmful bias may also result in regulatory fines and lawsuits for companies that don't take this risk seriously. Firms operating in Europe are well aware of the General Data Protection Regulation (GDPR), which states in section 9: "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or . . . data concerning a natural person's sex life or sexual orientation shall be prohibited."[4] A transgression could cost your business up to 4% of annual global revenue. In the US, the Consumer Financial Protection Bureau has already fined financial institutions over $300 million for discriminatory lending practices: charging minorities higher rates for loans.[5]

- **Revenue impact as values-driven customers go elsewhere.** Many of today's increasingly empowered customers consider company values when making a purchasing decision. Forrester questioned our ConsumerVoices Market Research Online Community and found that the majority of consumers would change their purchasing behavior if they discovered discrimination based on biased models. As one 27-year-old female consumer said: "I would definitely stop doing business with any company altogether if I found out that they discriminate against anyone! . . . If a business puts their product out in the public, they should be ready to cater to anyone interested and not deny them based on certain demographics!"[6]

Data Determines Discrimination

Here is a crash course on model-making with machine learning: data + algorithm = model. Machine learning models are therefore only as good as the data you use to train them. "Bias in, bias out" is just as true as the well-worn adage "garbage in, garbage out." An algorithm that is fed objectionable data will produce an objectionable model. For example, when Google trained a neural network using 3 million words from Google News, it identified strong relationships between words such as "man is to king as woman is to queen." However, it also picked up on sexist word relationships such as "man is to computer programmer as woman is to homemaker."[7]

Bias can be helpful, identifying important differences between customers that CI pros can use to drive customer acquisition, retention, and loyalty. Bias can also be quite harmful, blocking customers from accessing your products and services. Algorithms learn different types of bias in insidiously imperceptible ways (see Figure 1). CI pros should be aware of three types of learned bias in models:

› **Algorithmic bias due to incomplete training data.** When the training data used to teach a machine learning algorithm to perform its function does not accurately reflect the population the model will treat, the result is algorithmic bias. The term "algorithmic bias" is a misnomer because algorithms are inherently neutral — it is incomplete training data that creates biased models. The company FaceApp provides an example of algorithmic bias similar to the Google Photo example. It offers filters, like Snapchat, that alter a user's photos to appear younger, older, or of the opposite sex. Its "hot" filter was supposed to make photos seem more attractive, but instead it lightened the skin of dark-skinned users. When FaceApp created the "hot" filter, it didn't use a diverse enough set of images and the result was a racist app.[8]

› **Historical bias perpetuated by redundant encoding.** Another type of harmful bias occurs when a model uses a proxy for race, age, or another unethical discriminator to perpetuate historical bias or inequity in the training data. "Redundant encoding" occurs when a model includes one variable that is a proxy for another variable that should not be included. A classic example of this occurs when a lending decision should not consider gender, yet a bank asks if applicants are single parents. Since 82% of single parents in the US are female, the resulting model is actually using gender to make its decision.[9] Preventing this type of bias can be almost impossible. Solon Barocas, assistant professor of information science at Cornell University and author of "Big Data's Disparate Impact," states: "With multidimensional datasets it's almost certain your gender or race will be redundantly encoded."[10]

› **Real and exploitable differences between people.** Not all types of bias are harmful. In fact, CI pros are trying to teach models to discriminate between people, just not in a damaging way. Real and exploitable biases are the differences in preferences, behaviors, and propensities that algorithms pick up on to create models that drive differentiated experiences. Facebook's Lookalike Audiences provides a good example. It allows businesses to upload a list of their best customers; then Facebook targets users with similar attributes with ads for those brands. Ideally, this means that prospects who are more likely to convert will see your ads and others will not. So the model is discriminating between groups of people, but ostensibly not in a harmful way.[11]

FIGURE 1 Customer Insights Pros Need To Distinguish Between The Three Types Of Learned Bias

	Type of bias	Cause of bias	Solution	Responsibility
Harmful	Algorithmic bias	Training data is not representative.	Find or engineer independent and identically distributed (IID) training data.	Data scientists
	Human bias	Historical inequity captured in data	Modify training data or deployment engine to reflect a just outcome.	Line-of-business owner
Helpful	Useful bias	Actual behavioral and attitudinal differences between customers	Determine what types of bias it is OK to exploit.	Senior leadership

Code The Change You Want To See In The World

CI pros need to act as a company's first line of defense when it comes to preventing harmful bias in models. Due to challenges like redundant encoding, their task is not as straightforward as simply excluding sensitive attributes like race, gender, age, and sexual orientation from their training data. Even though most CI pros are not experts in business ethics, they can take a series of practical steps during the model-building and evaluation process to ensure their models aren't unintentionally discriminatory. Follow Forrester's framework for ensuring your models are FAIR (see Figure 2):

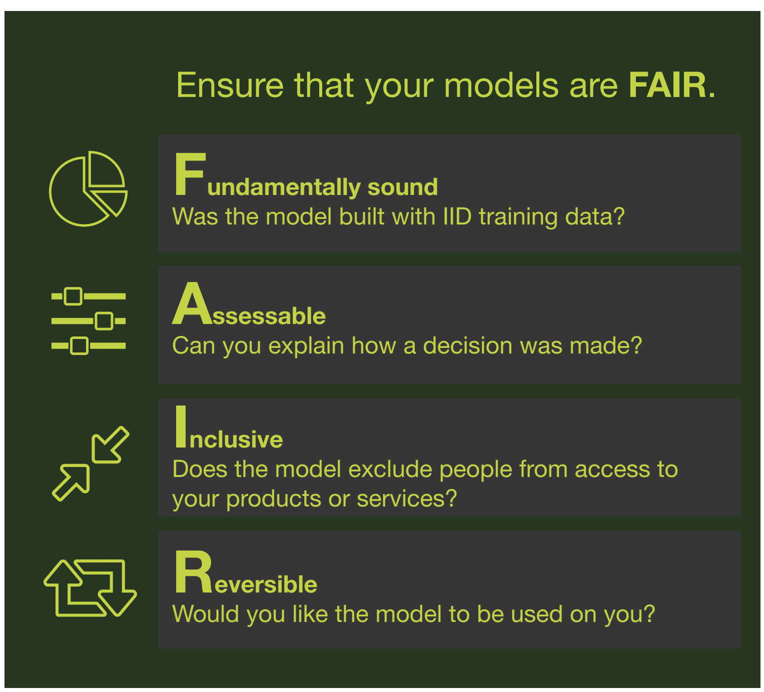
› **Fundamentally sound.** To prevent algorithmic bias in models, you need to adhere to fundamental principles of data mining by ensuring that your training data is representative of the population on which you plan to use the model. For example, CI pros building a facial recognition model for the general population of the US should strive to find a training data set that mirrors the racial makeup of the country.[12] Data scientists refer to this type of data as IID — independent and identically distributed. Since IID data is not always feasible to obtain, you may have to overrepresent certain groups. Otherwise, your models are likely to fall victim to the "tyranny of the majority," as Google Photo and FaceApp so notoriously discovered.

› **Assessable.** In many situations and in certain geographies (like Europe with the advent of the GDPR), CI pros will need to favor explainable machine learning models to assess how they make decisions.[13] This is because machine learning algorithms have varying degrees of transparency. For example, regression and classification models articulate themselves as equations or decision trees, clearly showing how they arrive at decisions and what variables they use to do so. On the other hand, artificial neural networks, the algorithms responsible for deep learning, are opaque by nature. Since CI pros employing neural networks cannot determine how the model makes its decisions, they will need to take extra care to ensure the decisions themselves aren't discriminatory. Fortunately, computer scientists at the University of Massachusetts Amherst have developed a software called Themis that tests models for causal bias to spot discrimination.[14]

› **Inclusive.** You must assess the impact of a machine learning model to determine whether it excludes customers from your products or services on the basis of race, age, gender, or other discriminatory factors. Often models that inherit historical bias exclude certain groups, like in the case of Amazon's same-day delivery service. If Amazon had assessed this model for inclusivity, it would have found that in Atlanta 96% of white residents were eligible for the service while only 41% of black residents were. While not necessarily illegal, this type of disparate impact is not something that aligned with Amazon's core values. So the company decided to roll out the services to all ZIP codes in the cities with same-day delivery service. Amazon changed its model to generate a more just outcome.

› **Reversible.** "Do unto others as you would have them do unto you." The Golden Rule is another tool for assessing the impact of your machine learning models. If the situation were reversed and you were in your customers' shoes, how would you feel about the way this model is treating you? Does it seem unjust, biased, or just plain greedy? If so, you probably need to reevaluate the model or even go back to the drawing board. Customer empathy will be a key weapon in the fight against harmful bias, and employees exhibiting a capacity for empathy should be recognized and rewarded for it.[15]

FIGURE 2 Customer Insights Pros Must Ensure That Their Models Are FAIR



IID: independent and identically distributed

Avoiding Harmful Bias Should Be An Organizational Priority

Protecting against harmful bias shouldn't just fall to CI pros and their data scientists — the leadership of your organization needs to make it a top priority. As more companies turn to machine learning to automate decision making, more examples of harmful bias are bound to emerge. Fortunately, there are several strategies for avoiding bias that fall outside the realm of models and algorithms. At the organizational level, firms need to:

› **Employ a diverse workforce.** The best way to ensure that your business practices aren't unintentionally discriminatory is to employ people of diverse ethnicities, backgrounds, and politics. By giving voice to multiple points of view, you protect your workplace from becoming an echo chamber. The tech industry in general suffers from an underrepresentation of female, black, and Hispanic employees. Pew Research recently found that 73% of Americans feel that discrimination against women is a major problem in the tech industry. The same survey found that 64% of black Americans and 50% of Hispanic Americans believe that discrimination against them is a major problem in the industry.[16] The folks who are most likely to spot bias are those who are most likely to feel its impact.

› **Listen to the opinions of customers and stakeholders.** Today it is essential to listen to your customers. Fortunately, it is also easier to do so than it used to be. Advances in text and speech analytics make it much easier to keep your thumb on the pulse of your customer experience.[17] Brands can mine customer feedback data from social media, online forums, internal feedback channels, or even customer service calls to determine when customers feel they're being treated unfairly. They can also solicit feedback by asking customers and stakeholders what they perceive to be the brand's core values — and whether the brand lives up to them.[18] And, of course, CI pros will want to rope in stakeholders from compliance, legal, and customer experience to ensure they properly deploy the model.

› **Continuously monitor results, as the impact may change over time.** One of the promises of AI is online learning — that models can continuously optimize by learning from new data in real time. This means that a model that disregards race in its decisions today may start to use race tomorrow if it becomes a better predictor of the model's objective. This is what happened with Tay, Microsoft's failed experiment at creating a chatbot to interact with users on Twitter. Within hours, the Twitterverse baited it into becoming what the Telegraph described as "a Hitler-loving sex robot."[19] To Microsoft's credit, it was monitoring Tay and shut it down within hours. Brands need to follow this example, monitoring the impact of their continuously learning models in real time and having a plan in place if the models start to become biased, offensive, or discriminatory.

What It Means

AI Will Push Ethics Into The Spotlight — For Firms And Society

Most of the examples of harmful bias today come from the tech giants because they are the most sophisticated when it comes to machine learning and artificial intelligence. However, we are at the beginning of an AI revolution, and as more companies adopt machine learning, we will see many more missteps. Businesses and society as a whole must engage in a discussion about ethics in AI — about when it is appropriate to treat different types of people differently using advanced machine learning. CI pros should help lead this discussion at their firms. Firms that take a proactive approach and tackle this head on will be better equipped to navigate these murky ethical waters than those that wait to react. In the next three to five years, Forrester predicts that:

› **Businesses will take the lead as regulators struggle to catch up.** Although regulators traditionally have a difficult time adapting their rules to advances in technology, values-driven consumers demand ethical practices today. For a time, companies will need to self-regulate by developing ethics-based centers of excellence tasked with answering big questions like "What are the ethics of our brand?" and "When is it OK to treat different customers differently?" Google recently did this, creating the DeepMind Ethics & Society unit "to help technologists put ethics into practice, and to help society anticipate and direct the impact of AI so that it works for the benefit of all."[20] Companies may also look to external professional associations for best practices. The Institute of Electrical and Electronics Engineers (IEEE) is currently creating a standard to help developers avoid algorithmic bias as part of its Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.[21]

› **Blame will fall on the shoulders of model integrators, not developers.** Since many brands will deploy machine learning models trained by a third party rather than building their own, there may be some initial confusion as to who to hold responsible when something goes wrong. That ambiguity won't last long. Just as FICO isn't held accountable when a consumer questions a bank's decision to deny them credit, so Amazon, Google, and other providers of trained machine learning models will not be held accountable for how other companies use their models. Instead the companies themselves, as the integrators of these models, will bear the consequences of unethical practices. To further protect themselves from regulatory exposure, AI vendors will likely include indemnity clauses in contracts, placing the onus of ensuring compliance firmly on the buyer.

› **Machine learning will serve as a moral mirror . . . and multiplier.** The proliferation of machine learning will expose human bias, acting as a "moral mirror," according to professor Shannon Vallor, who explores the ethical issues raised by emerging technologies at Santa Clara University. She points out that "models are not just pattern identifiers, but pattern amplifiers" in that they further codify bias by reinforcing it over and over again at scale. ProPublica recently found that the models many US states use to predict recidivism in the criminal justice system are much more likely to falsely label black defendants as high risk.[22] The result is harsher penalties and higher incarceration rates for black people, which in turn lead to higher rates of recidivism in an inexorable vicious cycle. We as a species stand at a crossroads where we can either repeat and reinforce the mistakes of the past or correct them and create a more just world for our future.

Engage With An Analyst

Gain greater confidence in your decisions by working with Forrester thought leaders to apply our research to your specific business and technology initiatives.

Analyst Inquiry

To help you put research into practice, connect with an analyst to discuss your questions in a 30-minute phone session — or opt for a response via email.

Learn more.

Analyst Advisory

Translate research into action by working with an analyst on a specific engagement in the form of custom strategy sessions, workshops, or speeches.

Learn more.

Webinar

Join our online sessions on the latest research affecting your business. Each call includes analyst Q&A and slides and is available on-demand.

Learn more.

Forrester's research apps for iOS and Android.
Stay ahead of your competition no matter where you are.

Supplemental Material

Survey Methodology

Forrester used its ConsumerVoices Market Research Online Community, Q3 2017 (US) as a source for this report. With Forrester's qualitative market research online community (MROC) of roughly 100 general-population consumers, we explored consumers' attitudes to and behaviors regarding brand content. Through a series of discussion board-style questions, we asked respondents to reflect on opinions and expectations of data algorithms and their impact on customer experiences.

Companies Interviewed For This Report

We would like to thank the individuals from the following companies who generously gave their time during the research for this report.

Capital Group

Enova Decisions

Epsilon

KPMG

Luminoso

Mighty AI

Rulai

simMachines

Teradata

Think Big Analytics

Voicera

Endnotes

[1] Source: @jackyalcine Twitter account, June 28, 2015 (https://twitter.com/jackyalcine/status/615329515909156865).

[2] Corporate reputation is a critical asset. Lululemon athletica discovered the importance of reputation after its founder made insensitive comments that resulted in lost sales and a severely damaged brand. See the Forrester report "Brand Resilience: Understanding Risk Managers' Key Role In Protecting Company Reputation."

[3] Source: David Ingold and Spencer Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?" Bloomberg, April 21, 2016 (https://www.bloomberg.com/graphics/2016-amazon-same-day/).

[4] Source: "Article 9 EU GDPR: 'Processing of special categories of personal data,'" PrivazyPlan (https://www.privacy-regulation.eu/en/9.htm).

[5] Source: Rick Rothacker and David Ingram, "Wells Fargo to pay $175 million in race discrimination probe," Reuters, July 12, 2012 (https://www.reuters.com/article/us-wells-lending-settlement/wells-fargo-to-pay-175-million-in-race-discrimination-probe-idUSBRE86B0V220120712).

Source: Ben Lane, "BancorpSouth fined $10.6 million for discriminatory lending, redlining," HousingWire, June 29, 2016 (https://www.housingwire.com/articles/37405-bancorpsouth-fined-106-million-for-discriminatory-lending-redlining).

Source: "CFPB and DOJ Order Ally to Pay $80 Million to Consumers Harmed by Discriminatory Auto Loan Pricing," Consumer Financial Protection Bureau press release, December 20, 2013 (https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-order-ally-to-pay-80-million-to-consumers-harmed-by-discriminatory-auto-loan-pricing/).

Source: "CFPB and DOJ Reach Resolution With Toyota Motor Credit To Address Loan Pricing Policies With Discriminatory Effects," Consumer Financial Protection Bureau press release, February 2, 2016 (https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-reach-resolution-with-toyota-motor-credit-to-address-loan-pricing-policies-with-discriminatory-effects/).

Source: Nathan Bomey, "JPMorgan pays $55M to settle mortgage discrimination lawsuit," USA Today, January 18, 2017 (https://www.usatoday.com/story/money/2017/01/18/us-accuses-jpmorgan-mortgage-discrimination-lawsuit/96710486/).

[6] Base: US online adults, 18+. Source: Forrester Data ConsumerVoices Market Research Online Community, Q3 2017 (US).

[7] Source: "How Vector Space Mathematics Reveals the Hidden Sexism in Language," MIT Technology Review, July 27, 2016 (https://www.technologyreview.com/s/602025/how-vector-space-mathematics-reveals-the-hidden-sexism-in-language/).

[8] FaceApp recently caused outrage again when it developed new racial filters to transform photos into Caucasian, Asian, Indian, or black categories. Source: Alex Hern, "FaceApp forced to pull 'racist' filters that allow 'digital blackface,'" The Guardian, August 10, 2017 (https://www.theguardian.com/technology/2017/aug/10/faceapp-forced-to-pull-racist-filters-digital-blackface).

[9] Source: Jennifer Wolf, "Single Parent Statistics," The Spruce, February 18, 2017 (https://www.thespruce.com/single-parent-census-data-2997668).

[10] In his paper "Big Data's Disparate Impact," Barocas explores the mechanisms by which biased machine learning models may lead to employment discrimination. Source: Solon Barocas, "Big Data's Disparate Impact," SSRN papers, September 30, 2016 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899).

For more information on the concept of redundant coding, see the paper "Fairness Through Awareness." Source: Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel, "Fairness Through Awareness," Cornell University Library, November 29, 2011 (https://arxiv.org/abs/1104.3913).

[11] Of course, bad actors may always use a machine learning model in harmful ways. Stanford University recently trained a model to guess a person's sexual orientation based on facial images, and it is surprisingly accurate. LGBTQ rights groups quickly denounced the research for many reasons; one of them being that homosexuality is punishable by death in 10 countries. Source: Sam Levin, "New AI can guess whether you're gay or straight from a photograph," The Guardian, September 8, 2017 (https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph).

[12] The United States Census Bureau provides a breakdown on its website. Source: United States Census Bureau (https://www.census.gov/quickfacts/fact/table/US/PST045216).

[13] GDPR section 22.

[14] Source: Janet Lathrop, "UMass Amherst Computer Scientists Develop New Technique to Measure Social Bias in Software," UMass Amherst press release, August 14, 2017 (https://www.umass.edu/newsoffice/article/umass-amherst-computer-scientists-develop).

[15] To build customer empathy, organizations should share customer insights, immerse employees in their customers' worlds, and incorporate empathy building into business processes. See the Forrester report "Customer Empathy: Three Essential Strategies For Cultivating It In Your Organization."

[16] Source: Kim Parker and Cary Funk, "Women are more concerned than men about gender discrimination in tech industry," Pew Research Center, October 10, 2017 (http://www.pewresearch.org/fact-tank/2017/10/10/women-are-more-concerned-than-men-about-gender-discrimination-in-tech-industry/).

[17] See the Forrester report "Deep Learning: The Start Of An AI Revolution For Customer Insights Professionals."

[18] Understanding what customers truly care about requires an "outside-in" approach that combines qualitative and quantitative data. See the Forrester report "The Values-Based Consumer."

[19] Source: Helena Horton, "Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours," The Telegraph, March 24, 2016 (http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/).

[20] Source: Verity Harding and Sean Legassick, "Why we launched DeepMind Ethics & Society," DeepMind blog, October 3, 2017 (https://deepmind.com/blog/why-we-launched-deepmind-ethics-society/).

[21] Source: "IEEE Announces Standards Project Addressing Algorithmic Bias Considerations," IEEE Standards Association press release, March 9, 2017 (http://standards.ieee.org/news/2017/ieee_p7003.html).

[22] Source: Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016 (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

FORRESTER®
CHALLENGE THINKING. LEAD CHANGE.

We work with business and technology leaders to develop customer-obsessed strategies that drive growth.

PRODUCTS AND SERVICES

- Core research and tools
- Data and analytics
- Peer collaboration
- Analyst engagement
- Consulting
- Events

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals
CMO
B2B Marketing
B2C Marketing
Customer Experience
› Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals
CIO
Application Development & Delivery
Enterprise Architecture
Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals
Analyst Relations

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866-367-7378, +1 617-613-5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

130023