**AI Discourse - SOCI 40133 Research Appendix**

**Jacy Reese Anthis**

**March 5, 2021**

# **1. MOTIVATION AND THEORY**

AI is beginning to reshape organizations across a range of industries (Fang et al. 2019; Amabile 2019; von Krogh 2018; Baum and Haveman 2020; Keding 2020; Pachidi et al. 2020; Raisch and Krakowski 2020). It has been labeled a “general purpose technology” with implications for organizations across many sectors, including its use to develop further technology (Cockburn, Henderson, and Stern 2018). Given this cross-context significance, scholars have begun to document the effects of AI on a range of organizational features, such as organizational control (Kellogg, Valentine, and Christin 2020), power asymmetries (Curchod et al. 2020), platforms (Gregory et al. 2020; Clough and Wu 2020), surveillance (Brayne 2017), revenue management (Lobel 2020), and the unequal distribution of organizational resources (Ahmed and Wahed 2020).

The primary goal of AI is, of course, the fulfilment of its intelligent capabilities, such as a predictive algorithm that makes accurate diagnoses based on medical imaging (Lebovitz, Lifshitz-Assaf, and Levina 2019). This does not always happen because of “algorithm aversion,” when people avoid incorporating algorithms into their work and decision-making routines (Dietvorst, Simmons, and Massey 2014; 2015; 2018; Kawaguchi 2020). In addition to this economic benefit of AI, there are a number of other values proposed by computer scientists, ethicists, and social scientists, particularly unbiasedness (Cowgill and Tucker 2019; Cowgill et al. 2020; Sunstein 2019; Schwemmer et al. 2020; Lambrecht and Tucker 2019; Obermeyer et al. 2019), fairness (Cowgill and Tucker 2019; Butterworth 2018; Cowgill, Dell’Acqua, and Matz 2020; Morse et al. 2020; Parkes and Vohra 2019), trustworthiness (Brundage et al. 2020; Marcus and Davis 2019; Kizilcec 2016; Glikson and Woolley 2020), and interpretability (Doshi-Velez and Kim 2017; Yu et al. 2020; Samek, Wiegand, and Müller 2017). Ultimately, Vamplew et al. (2018) categorize human-aligned artificial intelligence as a “multiobjective problem.”

I think the best contribution I can make to this conversation is a paper that models the composition, relationships, and evolutions of these goals. For the SOCI 40133, I hope to begin exploring AI discourse via the News on the Web (NOW) Corpus available on the RCC server. I am approaching this as computational grounded theory (Nelson 2020). This project is in the first step of this approach, pattern detection using computational exploratory analysis. After this class, I will embark on the second step, hypothesis refinement using human-conducted interpretive analysis, and the third step, pattern confirmation.

I envision the final output of this project as similar to the paper Augustine et al. 2019, “Constructing a Distant Future: Imaginaries in Geoengineering”, published in 2019 in *Academy of Management Journal*. That paper laid out five “social imaginaries” of geoengineering (e.g., launching particulates or mirrors into the atmosphere to reduce sunlight and cool the earth). The imaginaries were present since before 1990 up until 2016. For example, the initial framing was scientists treating the technology as a “technofix”, a logical step in humanity’s increasing control of the Earth, but then environmental critics brought in “human hubris” as a critical frame, highlighting humanity’s track record of harming Earth. This was a qualitative paper, but I hope to computationally track analogous frames in artificial intelligence discourse. See, for example, this figure from the paper:

Timeline

Description automatically generated

There are many other figures like this in management and organization papers that communicate the core substance of similar work. Some are grids like this one; others are diagrams with boxes and arrows.

Currently, it seems the most compelling theoretical foothold from which to do that is *frames*, the “schema of interpretation” popularized by the sociologist Erving Goffman (1974). Frames “organize experience and guide action, whether individual or collective” (Snow et al. 1986). I am also interested in *sensemaking*, which refers to “the social psychological and epistemological processes by which actors form an understanding of the situations they find themselves in” (Fiss and Hirsch 2005). The other most promising footholds are exploring these frames as “social imaginaries” (e.g., Augustine et al. 2019), “organizational goals” (e.g., Warner and Havens 1968), “values” (e.g., Hitlin and Piliavin 2004), concepts in a “conceptual space” (Hannan 2019), or ideas in a “field ideology” (Hehenberger, Mair, and Metz 2019). Each have pros and cons, but currently “frames” seem most tractable in top management journals. The word groups rendered by topic models have been interpreted as frames in previous work (Bail 2014; DiMaggio, Nag, and Blei 2013; Mohr and Bogdanov 2013; Fligstein, Stuart Brundage, and Schultz 2017).

All of these theories fit into broader literatures on organizational theory, culture, and cognition, which I am continuing to explore. I am eager for feedback on which of these are most promising, especially any for which the literature has falsifiable assertions that my research could critique. I have found very few, and this is my most important bottleneck at the moment.

# **2. EXPLORING THE CORPUS**

I am fairly convinced that most computational text analysis projects should start with the basics and only use more sophisticated methods where less sophisticated methods fail. So after gathering and cleaning the corpus for SOCI 40133, I was most interested in simple keyword counts. Figure 1 shows relative frequency (count-of-word \* 1,000,000 / total-words-in-year) of various keywords related to the economic and social dimensions of AI. I did not see evidence of the main trend I was looking for, an increase in mentions of “bias”, “ethic”, or “fair” over time or peaks corresponding with major AI events (e.g., AlphaGo defeating the world Go champion in 2016), though “bias” did increase from 2017 to 2020. I was surprised to see a huge increase in mentions of “revenue”, at least without a comparable increase in “profit”. If the general social trend were an increase in businesses with AI-related revenue or a near-potential for AI-related revenue, then I would expect both keywords to increase in relative frequency.

**Figure 1. Relative Frequency of AI Keywords Over Time**

Chart, line chart

Description automatically generated

Several of these keywords may also vary in usage over time, which is not captured by a simple count. For example, a 2012 article that mentions “bias” says, “Coupled with the data, though, we must have a much better understanding of decision making, which means extending knowledge about cognitive biases, about boundary work (scientists, citizens, and policymakers working together to weigh options on the basis not only of empirical evidence but also of values).” (“The Future of Big Data”, Pew Research Center). This refers to the cognitive biases of humans, which affect policy. A 2020 article uses “bias” a different way, “Current implementations of the software also perpetuate racial bias by misidentifying people of color far more frequently than white people.” This is a more machine-centric usage of “bias”.

These disparate trends suggest a need for clustering and topic models to aggregate words into coherent topics and detect connections between words as well as changes in topics over time. First, KMeans clustering using 3–50 clusters does not seem to perform well on the data. Unlike the multidimensional scaling of datasets in class, the AI news articles just look like a blob. The silhouette scores are very low.

A picture containing square

Description automatically generated

A picture containing icon

Description automatically generated

The top terms in each cluster had some coherence. For example, we see business, customers, digital, customer, and security in one cluster; another is china, Chinese, trump, trade, Huawei, government, and companies; another is facebook, people, game, just, apple, says, company. This gives me some sense of the frames through which AI are viewed (e.g. a business-oriented frame, a China-oriented frame, and a Silicon Valley-oriented frame), but the others are less interpretable.

I built a vanilla LDA topic model, toying with the parameters and getting the most coherent results with 10 topics. The outputs of the topic model are listed in Table 1.

**Table 1. Vanilla LDA Topic Model of AI Corpus**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** | **Topic\_3** | **Topic\_4** | **Topic\_5** | **Topic\_6** | **Topic\_7** | **Topic\_8** | **Topic\_9** |
| google | market | learn | china | health | model | business | say | say | security |
| app | company | human | say | patient | learn | company | people | india | say |
| apple | year | work | company | information | image | customer | work | global | government |
| device | growth | need | facebook | test | machine | service | year | industry | state |
| user | financial | people | chinese | study | network | cloud | game | country | law |
| camera | share | machine | people | disease | high | solution | think | development | public |
| amazon | stock | change | social | medical | base | platform | know | government | information |
| feature | investment | way | trump | say | fig | digital | want | year | africa |
| phone | report | science | coronavirus | research | process | market | come | innovation | national |
| iphone | increase | job | pandemic | help | algorithm | product | thing | sector | country |

In this run, Topic\_0 is Google, Amazon (not Apple), phones, devices, etc. Topic\_1 is business (e.g. markets, stocks, investments). Topic\_2 is less clear, but learning and science seem related. Topic\_3 is related to China, Trump, and the coronavirus. Topic\_4 is healthcare (e.g. health, patient, disease). Topic\_5 is perhaps more technical content (e.g. model, machine, network, algorithm). In another run, Topic\_0 has something to do with knowledge, gaming, and design. Topic\_1 is China and international trade. Topic\_2 and Topic\_3 are more generic, then Topic\_4 is autonomous cars and energy, and Topic\_5 is business and marketing. Again we see that this is in general a very business-oriented corpus.

These are interesting, but notably there is nothing on the “ethics” dimension, such as “bias” or “fairness”. Since this was a primary interest of mine, I looked for a Python package to conduct seeded LDA and found the ‘GuidedLDA’ package. It was a headache to install, and I ended up having to just copy the .py files into the LDA package. I used the following seed topics based on the vanilla LDA

['china','chinese','trump','global','government','state','world','country','president'],

['stock','financial','investment','growth','revenue','profit','company','market'],

['phone','iphone','apple','camera','device','app','facebook','google'],

['health','medical','covid','coronavirus','patient','patients','care','healthcare'],

['car','driving','autonomous','car','cars','vehicle','vehicles'],

['algorithm','model','learn','network','process','network','learn'],

And a seventh list of keywords for the topic I wanted to encourage:

['bias','race','fair','ethic','privacy','liberty','surveillance']]

Unfortunately, as shown in Table 2, the topics seemed less coherent than the vanilla LDA, and I did not get a stable “ethics” topic. I tried several parameters and did not get an “ethics” topic, but I am not very confident that the seeded LDA was set up properly. More detail is in the annotated code file.

**Table 2. Seeded LDA Topic Model of AI Corpus**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** | **Topic\_3** | **Topic\_4** | **Topic\_5** | **Topic\_6** | **Topic\_7** | **Topic\_8** | **Topic\_9** |
| said | company | google | health | technology | data | data | said | like | data |
| china | market | new | patients | new | learning | said | new | people | ai |
| government | business | apple | medical | ai | model | security | year | just | technology |
| world | year | app | covid | said | machine | facebook | university | time | business |
| chinese | growth | like | research | data | using | information | students | think | new |
| new | million | users | 19 | systems | used | media | world | world | digital |
| country | companies | amazon | care | energy | models | use | technology | way | learning |
| economic | services | company | healthcare | car | based | content | research | human | work |
| state | financial | phone | data | driving | deep | gt | team | going | need |
| global | data | video | patient | intelligence | neural | people | science | says | intelligence |

I was interested in how the topics evolved over time, which can also help us understand what the stable topics are within the corpus. So I focused on dynamic topic modeling (DTM), as detailed in the SOCI 40133 homework. The DTM took 36 hours to run on a single thread. I modified the vanilla LDA output code to create a series of tables for how each topic changed over the 11 years in the NOW corpus. This is too much output to list here, but Table 3 is an example of the most interesting topic, where—this is reading the tea leaves—the nature of “human”-“computer” interaction has shifted from “robots” (i.e., embodied AIs) to “models” (i.e., disembodied AIs) and from simply “thinking” about AI to actually “working” with AI.

**Table 3. Evolution of a Topic in the Dynamic Topic Model**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** | **2021** |
| human | human | human | human | human | human | human | human | learn | learn | learn |
| computer | computer | computer | computer | people | people | work | work | work | work | work |
| think | think | think | think | computer | work | people | learn | human | human | model |
| robot | say | say | people | work | learn | learn | say | say | machine | human |
| say | robot | learn | work | learn | say | say | people | people | say | machine |
| learn | learn | work | learn | think | think | machine | machine | machine | model | research |
| work | work | people | say | say | machine | think | need | need | research | science |
| people | people | robot | robot | robot | robot | robot | robot | research | people | study |
| science | science | science | machine | machine | computer | computer | think | university | need | patient |
| know | know | machine | science | way | way | way | way | science | science | people |
| year | year | know | way | year | year | research | research | way | university | need |
| machine | machine | year | know | science | research | need | university | think | study | university |
| way | way | way | year | know | know | university | computer | robot | way | test |
| university | university | university | university | thing | science | year | future | future | change | health |
| thing | thing | thing | thing | university | thing | science | science | job | health | image |
| research | research | research | research | research | university | know | year | change | patient | say |
| problem | problem | problem | come | need | need | thing | know | model | help | base |
| brain | brain | brain | need | come | future | future | create | help | job | high |
| need | need | come | problem | future | come | create | change | create | algorithm | process |
| come | come | need | brain | problem | create | come | job | computer | process | algorithm |

I also built *word2vec* and *doc2vec* embeddings of the AI NOW corpus. Both of these were easier with the multiprocessing capability of *genism.models.word2vec.Word2Vec()*. The embeddings were coherent. For example, similar to the king – man + woman = queen analogy, we can show that google + android – facebook = os/voice/ios. However, some other analogies do not work, and overall it was challenging to find relevant relationships between the word vectors.

Similarly, BERT classifiers and GPT-2 text generation showed coherent results. I trained a BERT classifier to differentiate articles published in 2015 and 2017, which are before and after AlphaGo’s victory over Lee Sedol in 2016, which was a big year for AI. Because I had some trouble differentiating the years with traditional machine learning classifiers, I chose a hopefully easy task as a basic check that the texts were in fact different. Ultimately I found a 68.6% accuracy using the Hugging Face BertForSequenceClassification.

For text generation, Tables 4 and 5 show some prompts and associated outputs. We see that, as expected (or at least hoped), the fine-tuned GPT-2 generates more AI-relevant text. I also asked GPT-3 what it thinks the future of artificial intelligence is, to which it gave a trivial or perhaps ingenious response.

**Table 4. Pre-Trained and Fine-Tuned GPT-2 Prompts and Output**

|  |  |  |
| --- | --- | --- |
| **Prompt** | **Pre-trained GPT-2** | **Fine-tuned GPT-2** |
| My favorite activity is | to play with your friends. I love to play with my friends. | to watch a video of a video game and then play it back to the computer. |
| The solution to the world's problems is | to create a world where people are free to choose their own path. | to create a new kind of economy that works for everyone. |

**Table 5. Pre-Trained GPT-3 Prompt and Output**

|  |  |
| --- | --- |
| **Prompt** | **Pre-trained GPT-3** |
| The future of artificial intelligence is | in question because some people say it is unpredictable and others say it is predictable. This is because right now some things are predictable and some things are unpredictable. |

# **3. FUTURE RESEARCH**

Through this qualitative and computational exploration of AI discourse, I have built some hypotheses that I think could be further developed or tested using these methods. Many of these would involve other corpuses of AI discourse, such as newspaper articles, analyst reports, press releases, earnings conference calls, or scholarly papers.

* H1 (Milestone Effects): AI milestones change the constitutions, frequencies, and relationships of frames.
  + H1a: After milestones, ethical frames grow in salience relative to performance frames.
  + H1b: After milestones, critical frames grow in salience relative to positive frames.
  + H1c: After milestones, the existential risk (e.g. Terminator and human extinction) frame grows in salience.
  + H1d: After milestones, there is more exploitation relative to exploration of technical AI strategies (e.g. Lasso, BERT, neural networks).
* H2 (Stakeholder Effects): The frames used by AI stakeholders vary based on how closely they are involved in AI.
  + H2a: Stakeholders more involved in AI (e.g. computer scientists, AI firms) are less likely to discuss ethics than those less involved (e.g. social scientists, activists).
  + H2b: Stakeholders more involved in AI discuss ethics in a more positive way than those less involved.
* H3 (Trade-Off Avoidance): Some frames are more likely to be discussed in the same document than others.
  + H3a: AI discussants avoid discussion of ethics and performance in the same document (i.e. the topic models have low intersection; the word embeddings have large cosine differences).
  + H3b: AI discussants avoid discussion of security and transparency in the same document.
* H4 (Organizational Outcomes): The use of AI performance frames, meaning a focus on AI that is economically efficient and productive, is associated with better organizational outcomes (e.g. investment, revenue) than the use of AI ethics frames, meaning a focus on AI that is fair and beneficial to society.
* H5 (The Power of Early Action): Frames that are common early in the history of AI have a disproportionate effect on later-stage AI discourse.

These hypotheses can be tested with corpus such as newspaper articles about AI (e.g., NOW, ProQuest), scholarly papers on AI in computer science, social science, and ethics (e.g., Scopus), analyst reports (e.g., Thomson One Investext), press releases (e.g., LexisNexis), mission statements (e.g., company websites), and earnings conference calls (e.g., Refinitiv). I tried to replicate my code on the ProQuest TDM server with a corpus of 410,915 newspaper articles from all of their news databases (e.g. ProQuest Historical Newspapers) from 1990 to 2021. Unfortunately, almost every step of data processing and analysis on the ProQuest server has to be recoded because their servers cannot handle large computational loads. For example, simply tokenizing 200,000 newspaper articles will cause the kernel to die. I have not yet had time to fully recode my preprocessing or analysis. It is a slow, arduous process.

My sense is that the best specific method for H1 (Milestone Effects) will be structural topic models (STM), which I can implement with the ‘stm’ R package, which comes highly recommended. However, it might be best just to use a series of vanilla LDA topic models or STMs without a time variable. This would allow a difference-in-differences estimate of stable snapshots in the corpus, such as all scholarly papers published in the 3-12 months[[1]](#footnote-1) after a milestone event compared to those published up to 9 months before a milestone event and more than 12 months after any previous milestone events. For a difference-in-differences estimate, we also need a comparison corpus. In the case of scholarly papers, that could be non-AI papers in the same discipline or papers on a similar but separate topic, such as blockchain. The biggest challenge with H1 will probably be finding a truly exogenous shock relative to a dependent variable of interest. For example, if my dataset is scholarly papers and my discontinuity is a new AI architecture (e.g. AlphaGo, which defeated world Go champions in 2016), then it may be endogenous within the scholarly discourse. I have spoken with a grad student who is using the 2012 ImageNet competition as an exogenous shock based on the reasoning that it was only a small group of scholars who developed that approach, such that it was exogenous to the vast majority of scholars. Even in that case, the grad student is having some trouble convincing economics-oriented scholars of the validity of their causal inference. In our discussion, it seemed the strongest case for causal inference in my context might be media and corporate analyst coverage of AI milestones, since the milestones are exogenous relative to the journalists and analysts, though media coverage analysis can sometimes be off-putting to management and organization journals because it is seen as superficial. It may thus be best to focus on corporate analysts (e.g. Benner and Ranganathan 2012).

For H2 (Stakeholder Effects), a series of topic models, separated by authorship (e.g. AI researchers, AI companies, AI-focused ethicists, AI-focused social scientists, journalists covering AI) seems straightforward. The biggest challenge here may be disentangling genuine differences in frames between the stakeholders and other discursive practices (e.g., diction, reading level, document formatting).

For H3 (Trade-Off Avoidance), and the more general research goal of mapping out AI frames and their relations, I will use topic models to measure the extent of topic overlap.

Aside from the hypotheses and general cartography of AI frames, I hope to conduct various robustness checks of the AI discourse models. For example, using text, can we see when AI winters happened (i.e. when interest in AI dropped)? Can we see when deep learning became a focus in 2012? While we should not expect the models to verify all of our less rigorous knowledge about AI discourse, the models should verify our most confident knowledge. Similarly, for H1a-H3b, there will inevitably be many different ways to construct the models, such as different inclusion criteria for the documents, and I will try to test many of them (e.g., with bootstrapping) to see if there is convergence.

Finally, I will continue reading through the corpus to get qualitative insight. I also hope to conduct interviews with various AI stakeholders to supplement the text analysis.

# **REFERENCES**

Ahmed, Nur, and Muntasir Wahed. 2020. “The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research.” *ArXiv:2010.15581 [Cs]*, October. http://arxiv.org/abs/2010.15581.

Amabile, Teresa. 2019. “GUIDEPOST: Creativity, Artificial Intelligence, and a World of Surprises Guidepost Letter for Academy of Management Discoveries.” *Academy of Management Discoveries*, February, amd.2019.0075. https://doi.org/10.5465/amd.2019.0075.

Augustine, Grace, Sara Soderstrom, Daniel Milner, and Klaus Weber. 2019. “Constructing a Distant Future: Imaginaries in Geoengineering.” *Academy of Management Journal* 62 (6): 1930–60. https://doi.org/10.5465/amj.2018.0059.

Bail, Christopher A. 2014. “The Cultural Environment: Measuring Culture with Big Data.” *Theory and Society* 43 (3–4): 465–82. https://doi.org/10.1007/s11186-014-9216-5.

Baum, Joel A. C., and Heather A. Haveman. 2020. “Editors’ Comments: The Future of Organizational Theory.” *Academy of Management Review* 45 (2): 268–72. https://doi.org/10.5465/amr.2020.0030.

Benner, Mary J., and Ram Ranganathan. 2012. “Offsetting Illegitimacy? How Pressures from Securities Analysts Influence Incumbents in the Face of New Technologies.” *Academy of Management Journal* 55 (1): 213–33. https://doi.org/10.5465/amj.2009.0530.

Brayne, Sarah. 2017. “Big Data Surveillance: The Case of Policing.” *American Sociological Review* 82 (5): 977–1008. https://doi.org/10.1177/0003122417725865.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 2020. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.” *ArXiv:2004.07213 [Cs]*, April. http://arxiv.org/abs/2004.07213.

Butterworth, Michael. 2018. “The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework.” *Computer Law & Security Review* 34 (2): 257–68. https://doi.org/10.1016/j.clsr.2018.01.004.

Clough, David R., and Andy Wu. 2020. “Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems.” *Academy of Management Review*, October, amr.2020.0222. https://doi.org/10.5465/amr.2020.0222.

Cockburn, Iain, Rebecca Henderson, and Scott Stern. 2018. “The Impact of Artificial Intelligence on Innovation.” *National Bureau of Economic Research*, March. https://doi.org/10.3386/w24449.

Cowgill, Bo, Fabrizio Dell’Acqua, Sam Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. “Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3615404.

Cowgill, Bo, Fabrizio Dell’Acqua, and Sandra Matz. 2020. “The Managerial Effects of Algorithmic Fairness Activism.” *AEA Papers and Proceedings* 110. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3523937.

Cowgill, Bo, and Catherine E. Tucker. 2019. “Economics, Fairness and Algorithmic Bias.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3361280.

Curchod, Corentin, Gerardo Patriotta, Laurie Cohen, and Nicolas Neysen. 2020. “Working for an Algorithm: Power Asymmetries and Agency in Online Work Settings.” *Administrative Science Quarterly* 65 (3): 644–76. https://doi.org/10.1177/0001839219867024.

Dietvorst, Berkeley J., Joseph Simmons, and Cade Massey. 2014. “Understanding Algorithm Aversion: Forecasters Erroneously Avoid Algorithms After Seeing Them Err.” *Academy of Management Proceedings* 2014 (1): 12227. https://doi.org/10.5465/ambpp.2014.12227abstract.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err.” *Journal of Experimental Psychology: General* 144 (1): 114–26. https://doi.org/10.1037/xge0000033.

———. 2018. “Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them.” *Management Science* 64 (3): 1155–70. https://doi.org/10.1287/mnsc.2016.2643.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. “Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding.” *Poetics* 41 (6): 570–606. https://doi.org/10.1016/j.poetic.2013.08.004.

Doshi-Velez, Finale, and Been Kim. 2017. “Towards A Rigorous Science of Interpretable Machine Learning.” *ArXiv:1702.08608 [Cs, Stat]*, March. http://arxiv.org/abs/1702.08608.

Fang, Christina, Chengwei Liu, Bo Cowgill, Jerker C. Denrell, Phanish Puranam, Zur Shapira, and Sidney G. Winter. 2019. “Machines vs Humans: How Can We Adapt Organizations to AI?” *Academy of Management Proceedings* 2019 (1): 12809. https://doi.org/10.5465/AMBPP.2019.12809symposium.

Fiss, Peer C., and Paul M. Hirsch. 2005. “The Discourse of Globalization: Framing and Sensemaking of an Emerging Concept.” *American Sociological Review* 70 (1): 29–52. https://doi.org/10.1177/000312240507000103.

Fligstein, Neil, Jonah Stuart Brundage, and Michael Schultz. 2017. “Seeing Like the Fed: Culture, Cognition, and Framing in the Failure to Anticipate the Financial Crisis of 2008.” *American Sociological Review* 82 (5): 879–909. https://doi.org/10.1177/0003122417728240.

Glikson, Ella, and Anita Williams Woolley. 2020. “Human Trust in Artificial Intelligence: Review of Empirical Research.” *Academy of Management Annals* 14 (2): 627–60. https://doi.org/10.5465/annals.2018.0057.

Goffman, Erving. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Boston: Northeastern University Press.

Gregory, Robert Wayne, Ola Henfridsson, Evgeny Kaganer, and Harris Kyriakou. 2020. “The Role of Artificial Intelligence and Data Network Effects for Creating User Value.” *Academy of Management Review*, March, amr.2019.0178. https://doi.org/10.5465/amr.2019.0178.

Hannan, Michael T. 2019. *Concepts and Categories: Foundations for Sociological and Cultural Analysis*. New York: Columbia University Press.

Hehenberger, Lisa, Johanna Mair, and Ashley Metz. 2019. “The Assembly of a Field Ideology: An Idea-Centric Perspective on Systemic Power in Impact Investing.” *Academy of Management Journal* 62 (6): 1672–1704. https://doi.org/10.5465/amj.2017.1402.

Hitlin, Steven, and Jane Allyn Piliavin. 2004. “Values: Reviving a Dormant Concept.” *Annual Review of Sociology* 30 (1): 359–93. https://doi.org/10.1146/annurev.soc.30.012703.110640.

Kawaguchi, Kohei. 2020. “When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business.” *Management Science*, June. https://doi.org/10.1287/mnsc.2020.3599.

Keding, Christoph. 2020. “Understanding the Interplay of Artificial Intelligence and Strategic Management: Four Decades of Research in Review.” *Management Review Quarterly*, February. https://doi.org/10.1007/s11301-020-00181-x.

Kellogg, Katherine C., Melissa A. Valentine, and Angéle Christin. 2020. “Algorithms at Work: The New Contested Terrain of Control.” *Academy of Management Annals* 14 (1): 366–410. https://doi.org/10.5465/annals.2018.0174.

Kizilcec, René F. 2016. “How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–95. San Jose California USA: ACM. https://doi.org/10.1145/2858036.2858402.

Krogh, Georg von. 2018. “Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing.” *Academy of Management Discoveries* 4 (4): 404–9. https://doi.org/10.5465/amd.2018.0084.

Lambrecht, Anja, and Catherine Tucker. 2019. “Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads.” *Management Science* 65 (7): 2966–81. https://doi.org/10.1287/mnsc.2018.3093.

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina. 2019. “To Incorporate or Not to Incorporate AI for Critical Judgments: The Importance of Ambiguity in Professionals’ Judgment Process.” *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3480593.

Lobel, Ilan. 2020. “Revenue Management and the Rise of the Algorithmic Economy.” *Management Science*, September, mnsc.2020.3712. https://doi.org/10.1287/mnsc.2020.3712.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. First edition. New York: Pantheon Books.

Mohr, John W., and Petko Bogdanov. 2013. “Introduction—Topic Models: What They Are and Why They Matter.” *Poetics* 41 (6): 545–69. https://doi.org/10.1016/j.poetic.2013.10.001.

Morse, Lily, Mike H. M. Teodorescu, Yazeed Awwad, and Gerald Kane. 2020. “A Framework for Fairer Machine Learning in Organizations.” *ArXiv:2009.04661 [Cs]*, September. http://arxiv.org/abs/2009.04661.

Nelson, Laura K. 2020. “Computational Grounded Theory: A Methodological Framework.” *Sociological Methods & Research* 49 (1): 3–42. https://doi.org/10.1177/0049124117729703.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Pachidi, Stella, Hans Berends, Samer Faraj, and Marleen Huysman. 2020. “Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing.” *Organization Science*, October, orsc.2020.1377. https://doi.org/10.1287/orsc.2020.1377.

Parkes, David C., and Rakesh V. Vohra. 2019. “Algorithmic and Economic Perspectives on Fairness.” *ArXiv:1909.05282 [Cs]*, September. https://arxiv.org/abs/1909.05282.

Raisch, Sebastian, and Sebastian Krakowski. 2020. “Artificial Intelligence and Management: The Automation-Augmentation Paradox.” *Academy of Management Review*, February, 2018.0072. https://doi.org/10.5465/2018.0072.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2017. “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.” *ArXiv:1708.08296 [Cs, Stat]*, August. http://arxiv.org/abs/1708.08296.

Schwemmer, Carsten, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. 2020. “Diagnosing Gender Bias in Image Recognition Systems.” *Socius: Sociological Research for a Dynamic World* 6 (January): 237802312096717. https://doi.org/10.1177/2378023120967171.

Snow, David, E Burke Rochford, Steven Worden, and Robert Benford. 1986. “Frame Alignment Processes, Micromobilization, and Movement Participation.” *American Sociological Review* 51 (4): 464–81. https://doi.org/10.2307/2095581.

Sunstein, Cass. 2019. “Algorithms, Correcting Biases.” *Social Research: An International Quarterly* 86 (2): 499–511.

Vamplew, Peter, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. 2018. “Human-Aligned Artificial Intelligence Is a Multiobjective Problem.” *Ethics and Information Technology* 20 (1): 27–40. https://doi.org/10.1007/s10676-017-9440-6.

Warner, W. Keith, and A. Eugene Havens. 1968. “Goal Displacement and the Intangibility of Organizational Goals.” *Administrative Science Quarterly* 12 (4): 539. https://doi.org/10.2307/2391532.

Yu, Haizi, Heinrich Taube, James A. Evans, and Lav R. Varshney. 2020. “Human Evaluation of Interpretability: The Case of AI-Generated Music Knowledge.” *ArXiv:2004.06894 [Cs]*, April. http://arxiv.org/abs/2004.06894.

1. Different timelines would make sense for different corpuses, since ideally we will capture the time of working on a document in which the authors are most influenced by outside events. For scholarly papers in computer science, this may be a few months before presentation at a conference. For scholarly papers in social science and ethics, this may be a year or two before publication in a journal. [↑](#footnote-ref-1)