1

HYBRIDIZATION: COMBINING ALGORITHMS, AUTOMATION, AND PEOPLE IN NEWSWORK

The Panama Papers was undoubtedly the biggest investigative news story of 2016. The Pulitzer prize–winning project built on a massive trove of 11.5 million leaked documents—more than 2.6 terabytes of data—concerning offshore companies and the powerful people behind them. Buried in those documents were scoops that led to the downfall of the prime ministers of Iceland and Pakistan, rocked the worlds of banking and sports, and exposed the shady business dealings of major companies such as Siemens.[1] The International Consortium of Investigative Journalists (ICIJ) coordinated close to 400 journalists working with the leaked documents as they produced more than 4,700 news articles based on the data.[2] The scale of the investigation simply dwarfed anything attempted up to that time. How did ICIJ and their partners pull it off? (Hint: there were no fancy artificially intelligent "robots" involved.)

The scale of the Panama Papers leak makes it almost unimaginable to consider *not* using heavy-duty computer power. But the real trick was to harness computing in a way that enabled the hundreds of collaborating investigative journalists to contribute their expertise and ability to contextually interpret what they were finding. If there were a mantra it would be, "Automate what computers do best, let people do the rest." On the one hand is the necessary task of converting the millions of leaked documents into digital text indexed in databases, something machines excel at using optical character recognition (OCR) algorithms. In the case of the Panama Papers ICIJ delegated the OCR process to about thirty machines operating in parallel in the cloud.[3] This allowed documents to be put into databases that could be searched according to lists of keywords. On the other hand are tasks related to figuring out what companies and people to search for in the first place, and then connecting those entities to find patterns that allude to improprieties, such as tax evasion. These are tasks that still fall heavily on the shoulders of knowledgeable people. ICIJ maintains a

collaboration platform that lets reporters post queries, documents, or comments to leverage the collective intelligence of partners.

The Panama Papers illustrates the power of combining human knowledge and expertise with the capabilities of machines to cope with an immense scale of data. Such complementarity between human and machine labor will continue to drive the evolution of newswork in the coming years. Wholesale substitution of reporting and editing jobs with automation is far less likely given the current state-of-the-art in technology. Meticulous estimates by economists suggest that only about 15 percent of reporters' time and 9 percent of editors' time is automatable using currently demonstrated technology.[4] Journalists are in fairly good shape in comparison to occupations like paralegals, who have an estimated 69 percent of their time that could be automated. Journalism jobs as a whole will be stable, though bits and pieces will fall prey to automation and algorithms.

Every job or workflow mixes different types of tasks with different susceptibilities to automation. Some tasks are highly skills-based, while others are contingent on knowing a set of specified rules, and still others rely on a store of knowledge or expertise that's built up over time.[5] An example of a skills-based task is keying in text from a digitized document so that it can be indexed. ICIJ could have trained people to do this work, but we would all be long gone by the time they finished. Algorithms have reached a high degree of reliability for this type of task and so offer a new opportunity for scaling up investigations. Entity recognition is an example of a rules-based task that involves marking a piece of text as referring to a particular corporation or person. This type of task reflects a higher level of cognition and interpretation but can be automated when the rules are well-established (that is, it's clear what constitutes an entity being labeled as a person rather than a corporation) and the data (in this case the output of the OCR process) feeding the task are reliable. Finally, knowledge-based tasks reflect those activities with high uncertainty, such as when data are vague and ambiguous. For an investigation like the Panama Papers, a knowledge-based task might be understanding the relationship between two entities in terms of the intents and obligations of those entities to each other and to the jurisdictions where they reside. Each macro-task will have a different composition of subtasks, some of which may be skills- or rules-based steps that are more amenable to automation. Knowledge-based tasks can be enhanced through complementary algorithms and user interfaces that allow an expert to work more quickly. Most workflows will not be entirely automated. Instead, different levels of automation will be involved at different stages of information production.

As technology advances, however, more and more artificial intelligence and machine-learning techniques will be introduced into investigations like the Panama Papers (as we'll see in Chapter 2). Algorithms are beginning to make headway in cognitive labor involving rule- and knowledge-based tasks, creating new possibilities to expand the scale and quality of investigations. Some of this technology will completely automate tasks, opening up time to reinvest in other activities. Other advances will be symbiotic with core human tasks and will, for instance, make finding entities and interpreting a web of relationships between banks, lawyers, shell companies, and certificate bearers easier and more comprehensive for the next Panama Papers. The challenge is to figure out how to weave algorithms and automation in with human capabilities. How should human and algorithm be blended together in order to expand the scale, scope, and quality of journalistic news production?

To understand how this blend may come about, it is important to delineate the capabilities and limitations of our two main actors. What are algorithms, and what is it exactly that they do? And, what is journalism, and what do journalists do? Answering these questions will pave the way toward designing the future of hybridized newswork.

What Do Algorithms Do?

An algorithm is a series of steps that is undertaken in order to solve a particular problem or to accomplish a defined outcome. A cooking recipe is an algorithm—albeit one that is (often) executed by a human. It consists of a set of inputs (ingredients) and outputs (the cooked dish) as well as instructions for transforming and combining raw ingredients into something appetizing. Here we are concerned with algorithms that run on digital computers and that transform and combine information in different ways—information recipes cooked by computer, if you will.

The singular term that describes algorithms that operate on information is "computing," formally defined as "the systematic study of algorithmic processes that describe and transform information."[6] A fundamental question of computing concerns what information processes can be effectively automated. Automation in turn has been defined as "a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator."[7] A related term is "artificial intelligence" (AI), which can be understood as a computer system "able to perform tasks normally requiring human intelligence."[8] The phrase "autonomous

technology” entails a version of full automation in which a system operates without human intervention, notwithstanding the human design and maintenance work that all designed systems require. Full autonomy is one extreme in a spectrum of options that blend humans and computers (see Figure 1.1).

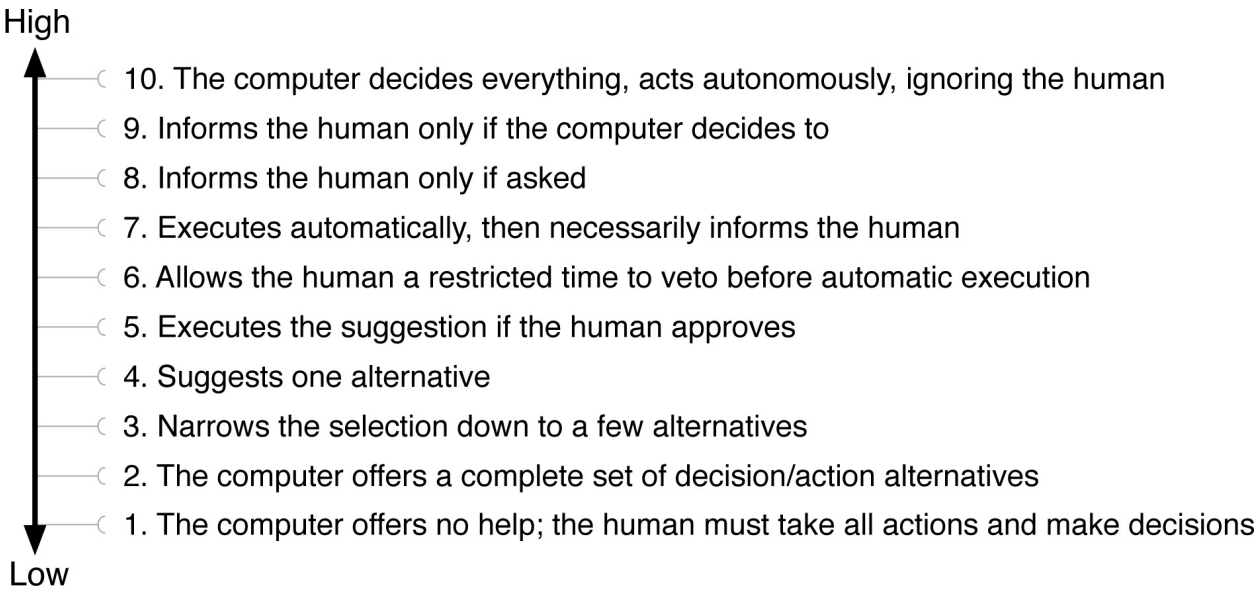

Figure 1.1. Levels of automation that blend more or less human and automated effort. *Source:* Figure derived from “A Model for Types and Levels of Human Interaction with Automation,” *IEEE Transactions on Systems, Man, and Cybernetics* 30, no. 3 (2001).

Much as machines and mechanization transformed the production of material objects in the nineteenth and twentieth centuries, computing is now transforming information work by offloading intellectual and cognitive labor to computers. This has been referred to as the “second machine age” because computers are now doing for mental work what machines did for physical work in the first machine age.[9] This cognitive labor encompasses computing tasks but also crosses into the terrain of AI to capture the idea of analytic information manipulation tasks typically associated with intelligence. Things start to get particularly interesting when algorithms enter into the evaluative phase of cognitive labor, in effect judging and *making decisions*. The quality of those decisions dictates how far we can push automation.

This has been a long time coming. As early as 1958 researchers at IBM described a program that could automatically extract an abstract from a research paper or news article.[10] In order to work, the system had to analyze each sentence and then judge how well it captured a key idea from the article. If it

was a representative snippet, then the algorithm would extract and add it to the summary. Fifty-five years later, in 2013, Yahoo! started using summarization technology in its news app to condense information from several news articles into a single briefing. The technology to analyze text by computer has been around for decades. But the automatic judgments needed to summarize an article have only recently reached a level of quality that allows the summaries to have actual value in the media marketplace.

Computer algorithms can do work in a few different ways. Some information tasks involve calculations of noncontroversial mathematical equations. Psychologists would call this an "intellective task," or a task with a demonstrably correct answer.[11] There are plenty of intellective tasks beneficial to information production processes. Digitization is a big one. Arrays of bits from audio or pixels from scanned documents—like the millions analyzed in the Panama Papers leak—need to be converted into recognizable words and symbols that can be further transformed and indexed in databases.[12]

But many tasks don't necessarily have a demonstrably correct answer and instead involve subjective judgment. Judgment tasks are politically interesting because they do not often have a correct answer. Instead, a preferred alternative is chosen based on facts as well as values, beliefs, and attitudes about the alternatives. The judgments that algorithms make are often baked in via explicit rules, definitions, or procedures that designers and coders articulate when creating the algorithm. Algorithms are neither neutral nor objective—though they will apply whatever value-laden rules they encode *consistently.* Machine-learning algorithms learn how to make decisions based on data. The algorithm is provided a set of observations about the world and learns how to make a judgment, such as a classification, by extracting patterns from those observations. The *New York Times* uses a machine-learned classifier to help it moderate comments on its site. Using data about which online comments have been flagged by a moderator as "toxic," an algorithm learns to classify future comments as "toxic" or "non-toxic."

The main value proposition of algorithms is their ability to make high-quality decisions, and to do so very quickly and at scale using automation. There are at least four fundamental judging decisions that algorithms make: prioritizing, classifying, associating, and filtering. Oftentimes these decisions are then composed into higher level information tasks. To take news article summarization as an example, such an algorithm must first *filter,* or select, a subset of representative sentences from an article and then *prioritize* them in

terms of importance to a user before presenting them as a summary. Other composite decisions are possible, too. The relevance of a search engine result could be considered a combination of an *association* decision between a search query and a result (that is, whether a particular website is related to a search term) and a *prioritization* decision that directs attention by communicating the magnitude of that association in a ranked list. All of these decisions rely on the calculation of analytic criteria, which themselves may be highly subjective, such as what defines a "representative" sentence or how one determines the "relevance" to a user for a ranking.

Prioritization decisions are perhaps some of the most crucial in the context of news media. A cousin to prioritization is optimization, which considers the top-priority item—the optimum along some dimension of priority. Given the limits of human attention, algorithms that can prioritize or optimize for the most interesting or informative content can select that content and present it first or give it privileged screen real estate so that it captures more attention. For instance, headline variations can be prioritized to pick the one that will optimize the click-through rate to an article. Designed into every prioritization decision are criteria that may be computed or derived and then used to sort items. These sorting criteria determine what gets pushed to the top and reflect editorial choices and value propositions that embed the design decisions of the algorithm's human creators.

Classification decisions also pervade newswork. For instance, organizations such as the Associated Press and the *New York Times* use algorithms to classify and standardize their vast content archives, allowing them to organize, store, transmit, or further process content in well-defined ways. Classification is highly political, involving decisions that range from what deserves to be a category to begin with to how categories are defined and operationalized quantitatively for computers.[13] Such algorithms can also be imbued with bias based on the input data they've been trained on. Human influence is woven into the process of defining, rating, and sampling the data to train the algorithm. Consider the toxic comment classifier again. The people who rate and grade comments to create training data end up having their biases built into the algorithm. Research has shown that men and women rate toxicity of comments in subtly different ways. So if men produce the majority of training data, as is the case for some commercially operational systems, then we can expect this bias to be refracted through the subsequent decisions the classifier makes.[14]

Association decisions denote relationships between entities. One example of

an associative relationship between two datasets is correlation, which indicates that as a value in one dataset increases or decreases, the corresponding number in another dataset also increases or decreases in step. Such a relationship implies a statistical connection between the two datasets. Of course, there are many other types of—and semantics for—associations that algorithms can help to identify, but they are always built on some criteria that define the association and a measure of similarity that dictates how precisely two things must match to be considered to have the association. For instance, in an investigation like the Panama Papers an association algorithm might be defined between two entities in order to link a person or company to another person or company in order to uncover or trace the flow of money. Such an association could be indicative of fraud, corruption, or a criminal scheme that is of interest to an investigative journalist.

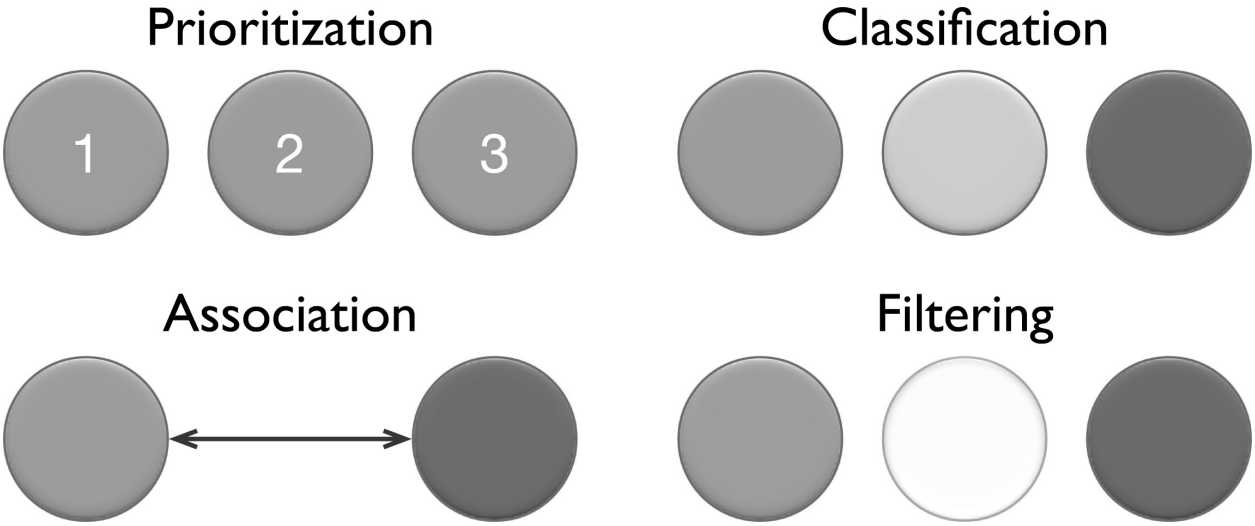

Figure 1.2. A schematic diagram of four fundamental information decisions. These can be composed into higher-level decisions such as summarization or relevance and are undergirded by calculations of analytic criteria.

Finally, algorithms can make decisions and take actions about what to selectively show, filter out, emphasize, or diminish, based on rules or criteria. Newsfeeds like Facebook's, news reading apps, recommendation widgets, and even news homepages make use of algorithms that dictate what to show or hide. This gets at a core function of what news organizations do: deciding what to publish or not publish. Filtering algorithms are increasingly used to help moderate social media by hiding offensive or uncivil posts that might disturb users. A news organization might deploy a toxic comment classifier by using the

toxicity rating as scored by the classification algorithm to filter from view those comments with a score above some preset threshold.

In summary, algorithms can do intellective tasks such as calculating things based on equations or transforming digital bits into words and symbols really fast, as well as make a range of decisions related to prioritization, classification, association, filtering, and compositions of these basic decisions (see Figure 1.2). Both calculating and decision-making algorithms have an immense potential to change the nature of information production. Yet automatic decisions are suffused with human judgments and values that undergird the various definitions and choices that constitute their design. The question of how far automation can penetrate into news and information production tasks depends on the types of decisions that need to be made in those tasks, and whether the algorithmic decisions made for a particular task are high enough quality to be accepted by end-users of that information.

What Is Journalism, and What Do Journalists Do?

A sound starting point for the function of journalism comes from sociologist Michael Schudson, who defines it as "the business or practice of regularly producing and disseminating information about contemporary affairs of public interest and importance."[15] In this sense journalism is about a relatively narrowly scoped production of information for society. But the concept of journalism can also be construed via an array of other activities and perspectives. To name just a few possibilities, journalism can be considered a practice, a profession, a business, an institution, a social field, or an ideology.[16] And the boundaries of what is and is not considered journalism are in constant flux: it is "a constantly shifting denotation applied differently depending on context."[17] Here I make use of the ideological view, which identifies shared beliefs in journalism about the importance of public service, objectivity, autonomy, immediacy, and ethics.[18] The ideology reflects a set of constitutive commitments—beliefs and codes—that journalists use to rationalize practices that are within the purview of journalism and that shape modes of thinking within the discipline.[19] Practitioners of journalism depict it as concerned with truth and verification, loyalty to the public, and independence and autonomy from those they cover, as well as being produced with an eye toward building community and fostering deliberative conversation.[20] Aspirational codes of practice, such as those from the Society for Professional Journalists, largely reflect and reinforce the ideological view.[21] Taking both the practices and the ideological commitments into account, I

consider journalism as a practice of news information and knowledge production that is filtered through a particular value system.

In the summer of 2009 I was a science reporting fellow at the *Sacramento Bee* newspaper, where I quickly got into a routine of calling sources for information, looking for datasets, reading scientific documents, and talking to editors as I scanned for my next story. As I made sense of the information collected, I would figure out an angle to frame the story and hook a reader's attention. Then there was the presentation of the story: perhaps I would just do a straightforward written article, but sometimes an intermingled data visualization or photograph would help illustrate a point. When it was all composed, it would of course get edited and finally published to the website and oftentimes in the printed newspaper the next day. My brief experience as a reporter made it easy to see the whole news production pipeline as information was transformed: from reporting and gathering of information to organizing and making sense of it, presenting and communicating it in a variety of media, and finally disseminating it to an intended audience.[22] Much of what journalists do on a day-to-day basis is taking raw observations of the world, including talking to sources or examining documents, and then transforming those observations into information and knowledge that they communicate to their audiences. In the process they make a variety of value-laden information judgments such as how to frame a story, what angle to focus on, and what is newsworthy—what is "of public interest and importance" in Schudson's words.

Journalists practice well-honed communication skills as they gather and then convey information. In so doing, they add a lot of value to information in transforming it from a "raw" state into a final easily consumed form of media. Information scientist Robert Taylor developed a helpful model for understanding how value is added during information production (value here is meant in the utilitarian sense rather than the ideological one).[23] Taylor suggests that as data is associated, related, and enriched, it becomes information. And as information is then validated, synthesized, and put into context, it becomes knowledge, which is in turn helpful for making decisions. As data becomes information and then knowledge, value is added. This is exactly what journalists do: increase the value of information for news consumers and for society.

Consider the Panama Papers investigation. The initial trove of leaked data contained thousands of documents for offshore companies: certificates of incorporation, copies of passports, lists of board members, and emails, among others. That data was transformed into information and given greater meaning

when journalists began to find connections between companies, transfers of money, and the people behind the operations. But it was only when those connections were validated and understood within the appropriate legal contexts that we could say the information had been transformed into knowledge, which in this case might be the certitude of malfeasance, for example, by a prime minister or major company. But this is just one specific example of value-adding in journalism. Taylor's model identifies at least four dimensions of value relevant to what journalists add to information in their daily practice: quality, usability, noise reduction, and adaptability.

Quality is of the utmost importance if the information and knowledge produced by journalists are going to be useful for making sound decisions in society. Quality can in turn be considered according to dimensions of accuracy (freedom from error), comprehensiveness (completeness of coverage), currency (up-to-date), reliability (consistent and dependable), and validity (well-grounded, justifiable, and logically correct). Journalists engage in quality control at many stages of information production in order to ensure that they produce trustworthy content. This involves everything from copy editing to remove errors from a text to triangulating sources when trying to verify and assess the reliability and validity of an image found on social media.

Journalists also add to the value of information by making it more usable. This could involve designing and presenting information in a way that is easy to consume on a user's device or that highlights the most relevant piece of information for a particular user. It could also entail making information more searchable or browsable to support goal- or non-goal-directed information access, or it could mean ordering or ranking content along some dimension of interest to make information easier to scan. The usability of information displays such as news apps, data visualizations, and video are increasingly important to news organizations seeking to enhance the value of their news offerings to end users. Even in terms of writing, the most routine journalistic activity, we can think about how a well-told story can enhance the usability of news information by making it more memorable, salient, and engaging.

Noise reduction is a result of decisions about inclusion and exclusion of information while maintaining focus and precision in the information that's delivered. In news production, noise reduction can involve clarifying and editing information about a major event to summarize what's known or curating and editing a collection of social media posts to focus on a topic of interest. Selection and filtering decisions often serve to help clarify information with respect to

quality, brevity, topicality, relevance, time spent, or really any other dimensions of editorial interest. Because of the paucity of human attention and immense competition for that attention, being able to reduce noise by focusing on the most important and relevant bits for news consumers is essential.

Adaptability captures the idea that information is used in particular contexts for making sense of particular problems or for making particular decisions. Two journalists could produce a story on exactly the same topic, such as corporate earnings, but one may present it for the sake of investors looking to make a trade decision, while another might cover it as an instance of a larger economic trend. News producers add value to information by aligning that information with how people will actually use it and by understanding what it is exactly that their audience hopes to glean from the content. For example, audience engagement editors routinely think about how content can be adapted or framed for different audiences so as to capture their attention.

At the end of the day, journalism is about ideology and values, and journalists are about increasing the value of information for their audiences. Together with commercial imperatives, the ideology of journalism drives journalists to add value to information across the news production pipeline whether by increasing quality, usability, and adaptability or by reducing noise. Beyond the strictly utilitarian, journalists produce value by helping people figure out where they fit in the world and by offering opportunities to identify with others or just find some entertainment.[24] At their best journalists do all of this in a responsible and ethical fashion that creates social value by supporting public understanding and democratic participation.

Can Algorithms Do Journalism?

As I've just outlined, journalism describes a set of practices for news information and knowledge production that are aligned with a particular journalistic ideology. Can that ideology be reflected in news production algorithms?

Yes!

At their core both journalism and computing share a focus on transforming and adding value to information. Computing approaches information from an algorithmic perspective whereas journalism focuses on information production practices that are informed by particular ideological commitments. Because algorithms can act to produce information and knowledge, and do so in light of values that are imbued through their design, algorithms can indeed do

journalism. Of course they need not. Alternative values, such as those of noneditorial stakeholders in news media, people dominant in other fields or in society at large, or end users interactively tweaking and tuning, may infuse algorithms instead.[25]

With this as background, I define computational journalism as *information and knowledge production with, by, and about algorithms that embraces journalistic values*. While others, including myself, have proffered other characterizations in the past,[26] here I wish to emphasize that computational journalism involves exploring the relationship between the underlying values of journalism and the ways in which algorithms are both designed and incorporated into news information production practices. Given the affordances of computation itself, computational journalism will not just mimic the value propositions of journalism (though it could), but will rather blend the ideology of journalism with the inherent affordances and values of computing, including, for example, an emphasis on scale, speed, and abstract problem-solving while relying on a quantified version of reality.[27] This book focuses on "computational journalism" rather than related terms such as "data journalism," "computer-assisted reporting," "interactive journalism," "algorithmic journalism," or "automated journalism" because "computational journalism" hews most closely to the idea of algorithmic information production that incorporates journalistic values.[28]

Technology has coevolved with the tasks of journalism throughout history, changing both the pace and structure of work, while shaping the content and industry too.[29] Each technology has its own values that may subtly permeate how information meets the public. These embedded values may offer opportunities for continuity in professional practices, but may just as well offer affordances that create tension with traditional journalism values.[30] As a technically oriented domain, however, computational journalism need not adopt the technologies others create and imbued with their own values.[31] With a distinct focus on designing "practices or services built around computational tools in the service of journalistic ends,"[32] the field is oriented toward designing and building technologies and algorithms to reflect the journalistic ideology. A stalwart computational journalist might declare a need for independence from the biases and values inherent in tools built by nonjournalists. Algorithm design will become the new way of exercising journalism so that the ethical responsibilities of the profession are met in the implementation and expression of journalistic values via code.[33]

A side effect of deliberately designing value-laden technology is that in order to articulate the set of steps in an algorithm, designers should be able to explicate and justify those steps in advance instead of after the fact (as is typical of justifications of journalistic activity[34]). For example, in order for a data-mining algorithm to detect a story lead in a large dataset, it must embody some clearly articulated and mathematically precise notion of "newsworthiness." In effect, practicing journalism using algorithms prompts an explicit consideration of an information selection process and its justification ahead of time. But explication of the factors built into a system has the benefit of allowing for discussion, debate, and deliberate adjustment.[35] Algorithms can and do express the values embodied in their design, and so by adopting more cognizant and reflective practices, value-sensitive designers can develop algorithms intended to operate within the ideological framework of journalism.[36] Value-driven modes of design thinking can help create technologies and algorithms that reflect journalistic priorities—something news organizations should consider if they want to ensure their values are present in the algorithms that drive the future of the media.

Can Algorithms Do What Journalists Do?

Algorithms can produce information while enacting the values of journalism and therefore they can do journalism. But can they execute a range of tasks and practices that are recognizable or analogous to what journalists do? Not entirely. Just as journalists add value to data and information, so too can algorithms. But they are oftentimes still limited in their capacities to do so.

Let's take the value-added journalistic practice of adaptability and consider a specific algorithmic application of adaptability: providing personalized content recommendations. Based on the topical interests of a user, an algorithm can *associate* content with an individual based on a *classification* of the content's topic. Using the magnitude of that association, the algorithm can then *prioritize* and *filter* content, surfacing a set of personalized recommendations for each person. The quality of algorithms to add this type of value to information is quite advanced and allows them to operate in the high end of the autonomy spectrum (see Figure 1.1). But what if we want to design this recommendation algorithm so that it balances personal interests with the importance of content to a local community deliberation, thereby better fulfilling the ideological goal of building community awareness. How should an algorithm know that a story should be shown to everyone regardless of their personal interest? Algorithms are not yet up to the task of calculating something like the social, political, economic, or

deliberative significance of a piece of content. Assessing those kinds of factors is better left to a person who has deep contextual knowledge of the community and an understanding of the myriad routes through which the news item could impact an issue in that community. So while a content recommendation algorithm can operate autonomously, in some situations we might still need it to be augmented with human capabilities if we want it to reach its full journalistic potential.

The effective and ethical design of news production algorithms will entail partitioning information and knowledge tasks: which should a person be making, and which can be reliably delegated to an algorithm?[37] To decide this, we need to understand both the decision-making capabilities of algorithms and the mental acuities and advantages of humans. The frontier of what types of cognitive labor algorithms are capable of is constantly shifting, but in *The New Division of Labor* Frank Levy and Richard Murnane posit that there are two key domains where humans have an edge over computers, and may still for some time: complex communication and expert thinking.[38]

Complex communication involves the two-way exchange of information and includes activities such as listening, negotiating, persuading, and explaining across both verbal and nonverbal channels. This sounds a lot like reporting, the bread and butter of journalistic information gathering, but it also includes tasks such as interpreting information to present an angle in a written news story, adapting information for different storytelling technologies and media, incorporating the current zeitgeist and public agenda, and putting information into context to meet particular audience needs.[39] Because journalism is so reliant on gathering information from people, complex communication also encompasses the social intelligence needed to engage empathetically or emotionally in a range of situations. Collecting information can involve undertaking difficult interviews with sources unmotivated to share information, perhaps even deceptive or antagonistic in their interactions. Developing trust with sources so they feel comfortable sharing sensitive information that might paint themselves or their organizations in a negative light is no easy task. Negotiating for information involves a push and pull of knowing when and how to convince an individual or organization to open up. And asking a source the “right” questions involves intent listening and reacting in the moment to a conversation that may be unfolding in unpredictable ways. While not highly automatable, many complex communication tasks can still be enhanced by technologies that complement human practices, such as a voice recorder that offloads a memory burden from a reporter or a spell-checker that improves the

quality of copy a reporter produces.

Expert thinking, on the other hand, involves the ability to solve problems effectively using domain knowledge. Some of this knowledge may be tacit or difficult to express formally. Complex problems often require some out-of-the-box thinking to know what's working and what's not, and to apply metacognition to identify when a problem-solving strategy needs to be switched out because it no longer seems promising. Oftentimes expert thinkers apply pattern matching based on detailed knowledge of a domain, using analogies to intuitively map new problems into more familiar ones. While not every task in news production entails expert thinking, investigative journalism of the Panama Papers variety certainly does. Investigation can include analyzing documents, data, and other sources for relationships and associations that may not be known ahead of time and whose significance and verity may become clear only through the interpretation of an expert with deep domain knowledge.

Human abilities in complex communication and expert thinking exhibit particular value in nonroutine situations. While algorithms excel at encoding and executing rule-based tasks, consistently and tirelessly responding to expected events at great speed, performing repetitive actions reliably, and detecting anticipated patterns, their downside is their inflexibility and inability to cope with unanticipated scenarios.[40] This is a key weakness in applying algorithms to newswork. Algorithms also lack the human capacity for creativity. By combining many different pieces, they may at times appear to produce novelty, but they are currently extremely limited in their ability to operate in new situations or conceptual spaces.[41] Rare is the algorithm that can surprise and delight in entirely unanticipated ways. The inflexibility limitation extends to complex communication abilities, too. As storytelling formats, technologies, and modes of interaction with audiences evolve, human adaptation of content presentation will be essential. Furthermore, algorithms and, in particular machine-learning approaches, are simply unsuitable in some scenarios, particularly those involving complex chains of reasoning, diverse background knowledge, and common sense, as well as those in which there isn't at least a modicum of tolerance for statistical error.[42]

Yet there may still be ways to transform some aspects of expert thinking and complex communication into more structured, systematized, and routinized tasks in which algorithms can be brought to bear. One approach to designing the frontier of what algorithms are able to accomplish for news production is to adopt a computational thinking mindset. Computational thinking is defined as

"the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent."[43] It is important to emphasize that computational thinking is *not* about getting people (journalists in this case) to think more like a computer. It's also not about writing computer programs per se. It's really a way of thinking about how to best use a computer to solve a problem, oftentimes at scale. In a way computational thinking is a reflection of the value system of computer scientists, who are trained to formulate and solve problems using computers. Computational thinkers will ultimately be more effective at exploiting the capabilities of automation when they see ways to structure and routinize processes to be executed by computer. While perhaps not a universally necessary skill for journalists, computational thinking capabilities will be essential for those wishing to be at the forefront of future algorithmic news production processes.[44]

A key tenet of computational thinking is abstraction. Seeing a specific problem and recognizing that it is an instance of a more general problem allows computational thinkers to recognize opportunities for applying computers to solve the larger-scale general problem. The algorithm can then solve the problem over and over again, thereby allowing for the benefits of computational scale to be realized. Abstraction is evident in the various chart-, map-, meme-, or quiz-making tools that have proliferated at news organizations such as *Vox*, *Quartz*, and the *New York Times*.[45] Each tool creates an abstract template that encodes a specific and particular form and style of content. For instance, the Mr. Chartmaker tool from the *New York Times* streamlines chart creation, while also making the output charts look more consistent.[46] Systematizing the authoring process and outputs allows news organizations to create more content, more quickly on deadline, and with less skilled content creators.

An important aspect of abstraction is parameterization, a process for creating procedures that can apply to a range of cases or contexts via parameter substitutions. Let's look at parameterization in terms of an analog algorithm for baking a cake. Suppose one of the ingredients called for by the recipe is eggs, but we want to adapt the recipe to make a vegan cake. What are eggs to the recipe really? Eggs are something to keep the other ingredients adhered together in the batter, a binding agent. For a vegan version of the recipe we can use a different binding agent, such as ground flax meal. By abstracting the binding agent as a parameter for the recipe we can use a parameter of "eggs" for the regular cake and a parameter of "ground flax meal" for the vegan cake.

Parameters enable a combinatorial explosion of options for abstracted algorithms, allowing them to achieve many different outcomes and suit a much wider range of contexts.

Modeling is a process closely related to abstraction. Models encode simplified representations of the world to describe objects and their relationships. Models can also be statistical in nature, articulating mathematical associations between variables of interest and allowing for prediction based on new data. Modeling is largely an editorial process of systematically deciding what is included, excluded, or emphasized by a particular representation of the world. For example, user models are often used by news organizations to articulate an abstracted view of their audience. Dimensions in the model might include a user's age, interests, income level, geography, occupation, education level, marital status, or other factors. Although such a model is a limited approximation to any given individual visiting a site, it does enable some useful outcomes. For instance, article recommendations can be made systematically according to interests, and advertisements can be targeted based on geography.

The final component of computational thinking is decomposition. Many processes or tasks entailed in producing the news are composed of smaller actions. Decomposition is about pulling apart the steps of a process to get at those smaller actions and tasks. Upon examining a big gnarly process and breaking it into simpler subtasks, the computational thinker will be able to identify which of those smaller tasks might be reliably solved by a computer. Decomposition provides a lens for process re-engineering using automation and algorithms. Of course, some subtasks may still need human attention and thus can't be automated. But by disaggregating a process, we can see what components are suited for a machine and what components are suited for a person, and then recombine these subtasks to more efficiently solve the larger problem.

Whether algorithms can do what journalists do is a moving target that will ultimately depend on whether the practices of journalists can be abstracted, parameterized, modeled, and decomposed in a manner that enables designers to see how to systematize processes and insert automation as a means of substituting or complementing human activity in constructive ways. The pieces of the work that can be routinized may be automated (such as OCR or entity recognition in something like the Panama Papers investigation), but in very few cases does such routine work constitute the entirety of a job in journalism. Human tasks will still account for the nonroutine exigencies of covering news

events that emerge from a messy and unpredictable world. Despite strong routines in journalistic work, there are still creative and improvisational scenarios demanded by news events that break with expectations.[47] Not all journalistic decision-making will be amenable to algorithms: this includes ethical judgments in particular, but really any judgments for which quantifications are not available or feasible. Still, computational thinking will help point the way to where algorithms can be effectively deployed.

Toward Hybrid Journalism

I've argued in the previous sections that algorithms can do journalism, and that they're advancing onto the turf of what journalists do, but that there are fundamental tasks of complex communication and expert thinking that will be complemented rather than replaced by algorithms. The future of computational journalism is in finding ways to harness computational thinking skills to invent new methods for combining human and computer capabilities that reinforce each other and allow the appropriate delegation of work. Stated more simply: How do we design and build an effective hybrid journalism? The role of algorithms is unavoidable in the future of journalism, but so too is the role of people.

Designing hybrid journalism won't be easy; there's no cookbook, no algorithm here. Yet production processes will need to be reinvented to take full advantage of technical capabilities. This reinvention is complicated by a sociotechnical gap: the divide between what we know we need to support some sophisticated human activity (such as complex communication) and what we know can feasibly be supported.[48] The allocation of tasks between humans and computers will emerge from a design process that entails iterative prototyping, development, and testing by a variety of entrepreneurs, established organizations, and research labs.[49] Processes will need to be carefully designed to take into account what computers and humans do well and to make the outputs of what each produces seamless, usable, and interoperable with what the other produces. Difficult design questions such as how to cope with nuance and uncertainty in algorithmically driven journalism will need to be grappled with and surmounted. Innovation will be needed to re-engineer processes and practices around information production while ensuring those new processes meet stakeholder expectations, including the quality and accuracy of content for audiences, the autonomy and satisfaction of journalists, and the bottom line of news organizations. Ideally hybrid workflows will both lower costs and enable an entirely new echelon of breadth, comprehensiveness, adaptability, speed, and

quality of content, which will unlock new possibilities for original, unique, and exclusive material that will be valuable to organizations seeking to compete in a largely commodity information market.[50]

Journalism studies scholars have begun to set the stage for this future by examining the decomposability of journalistic work through the theoretical lens of actor network theory (ANT). ANT considers behavior as emerging from an assemblage of humans (actors), objects and technologies (actants), and their relationships. For computational journalism it is essential to recognize the range of the actors and actants—both human and nonhuman, and inside or outside the newsroom—and examine the associations they engage in as news information is produced.[51] We must ask not only "who" does journalism, but also "what" does journalism, and that "what" includes technical artifacts and algorithms.[52] Understanding how to design assemblages of actors and actants to organize the work of producing news is a fundamental question for the future of computational journalism. As economists Erik Brynjolfsson and Tom McAfee argue, "Effective uses of the new technologies of the second machine age almost invariably require changes in the organization of work."[53] The relevant question here is then: Which actors and actants need to be put together, and in what ways, in order to accomplish some particular information or knowledge transformation task?

Studies of crowdsourcing offer instructive lessons. Crowdsourcing is fundamentally concerned with how tasks are accomplished in a distributed fashion by a set of people connected via a computer network. It often involves decomposing tasks into smaller tasks that are then completed and recomposed or synthesized into a final work output. Various news production tasks are already being reimagined so they can feasibly be carried out with crowdsourcing, including tasks such as copy editing, article writing, and reporting and information gathering.[54] Crowdsourcing has also been studied with respect to broader processes in news production, such as using crowds to check documents during investigations, verify locations and context for social media content, serve as a source of distributed knowledge, and co-develop ideas or brainstorm topics.[55] Examining how work is decomposed for crowdsourcing workflows suggests ways, as well as challenges, for the completion of work by assemblages of novices, experts, and algorithms.[56]

Much research remains to be done in developing a design science to grapple with the challenges of creating feasible human-computer workflows for news information and knowledge production. For instance, workflow design must

ensure that tasks can be decomposed and also recomposed without loss of information, and while ensuring a high-quality output on par with legacy modes of production. The human workers in hybrid workflows typically serve to maintain quality, either by preprocessing data fed into algorithms or by postprocessing algorithmic results.[57] Particularly in the news domain, workflows should not be restrictive or rigid, given that this could inhibit the ability to deal with contingencies present in work that is complex and unpredictable, has dynamic interdependencies, or is heavily time-constrained.[58] In order for some subtasks to be automated, fragments of those workflows will need to be parameterized, while leaving open opportunities for collaborating humans to adapt the algorithm to unforeseen circumstances. Humans integrated into the workflow may help to lubricate the automation, allowing it to flex and adapt as needed.

The humble chat system offers an important interface between people and algorithms working together. For many years now, chat systems have allowed groups of people to organize workflows without necessarily making all roles or information explicit.[59] The muddiness of an unstructured chat interface allows people to coordinate behavior in flexible ways. This may to some extent explain the rise of the use of the Slack messaging platform within newsrooms.[60] Not only does it support flexible and relatively unstructured coordination, but it also enables the integration of automated scripts or bots that can interject or help when tasks can sensibly be delegated to a machine. An intermediary platform like Slack functions as a glue for coordinating human and automated workers as workflows evolve.

To fully realize hybrid systems, there are key challenges that need to be resolved relating to task dependency. A Microsoft Research project ran into dependency issues when it developed a crowdsourcing process to write articles about local events.[61] Work was decomposed across four roles: reporter, curator, writer, and workforce manager, whose tasks were coordinated via email or Twitter. One of the difficulties for workers in the reporter role was that they lacked context and found it difficult to know how to cover an event without appropriate background knowledge or indeed training in how to approach individuals for interview purposes. “Breaking down the task into component pieces, as well as distributing it to several people, created fragmentation that led to context loss,” the researchers wrote.[62] In effect, for this model of task decomposition to function effectively, work must be broken up into small and independent pieces of effort in which dependencies between pieces are

understood and managed. Otherwise the decomposition and recomposition of work can end up introducing more overhead and trouble than they're worth.

Another factor to consider is the economic viability, or total cost and effort, involved in a hybrid workflow. This will depend heavily on both the complexity of the task and its prevalence. The greater the complexity of the task, the higher the fixed costs of designing and programming an automated solution and the higher the costs of recomposition of work from subtasks. The costs of the initial programming of an automated solution can of course be amortized depending on the prevalence of the task, thus modulating the cost per unit output. Ideally, the additional costs associated with recomposing work outputs from automation and from other human actors are less than simply having an individual undertake the macrotask on his or her own.

Designing hybrid systems demands a degree of creativity to understand how human and machine work can amplify each other. Steven Rich, a database editor at the *Washington Post,* provides a good illustration of workflow innovation by journalists who code. In the course of developing the *Post*'s Fatal Police Shooting Database he found himself needing to file around fifteen Freedom of Information (FOI) requests every week in order to get the necessary details from individual police jurisdictions. But he realized he could delegate this task to machines, so he programmed a script to feed information from a database into a form letter that could be automatically sent as a FOI request. This allowed him to offload the routine aspect of the work to a computer program, saving him from having to perform a repetitive task every week.[63] Of course, after the records requests were fulfilled, a person still had to read through the documents and key in and validate the data.

Trial and error will be required as different alternatives are prototyped and tested. The design of workflows themselves is, however, likely to remain a human endeavor. In instances where algorithms do substitute directly for humans in completing subtasks, interesting questions about management will arise: human workers will need to flag exceptions as well as have agency to stop and start processes in light of evolving conditions.[64] One could even imagine cases in which an algorithm becomes manager, delegating subtasks to human workers and managing the reconstitution of the work. Who, then, will have the authority to override an automated component, and under what circumstances?

Difficult questions remain as responsible and ethical approaches to hybrid news production emerge. But as I'll show again and again throughout this book, the adoption of hybrid workflows in practice is already well underway. As of

2018, roughly a quarter of Bloomberg News content already incorporates some degree of automation, a proportion that will only grow as news producers get better at blending human and computer capabilities.[65]

In this chapter I've articulated the central value proposition of algorithms: more effective and efficient decision-making. They calculate. They judge. They offload cognitive labor from people and make information jobs easier. Computational journalism, in turn, is the study of information production using algorithms operating within the value system of journalism. As the frontiers of what is possible to accomplish with automation, algorithms, and hybrid systems continue to expand, human journalists will still have a lot to add when it comes to complex communication, expert thinking, and ethical judgment—essential elements at the core of journalism that will resist the application of algorithms. In the following chapters we'll see these ideas play out in various different contexts: data mining (Chapter 2), automated writing systems (Chapter 3), newsbots (Chapter 4), and distribution algorithms (Chapter 5). Clever hybridization of algorithmic and editorial thinking will be the key throughout.