# Behavioral Experiments with Social Algorithms:

# an information theoretic approach to input-output conversions

Martin Hilbert[*#], Billy Liu[+], Jonathan Luu[+], Joel Fischbein[+]

* Department of Communication, University of California, Davis
+ Computer Science, University of California, Davis
# corresponding author, hilbert@ucdavis.edu

## ABSTRACT

While traditional computer-mediated communication happened through transparent, passive, and neutral channels, today's communication channels are obscure, proactive, and distorted. Social algorithms, guided by a socio-technological codependency, often bias communication, usually in pursuit of some third-party goal of commercial or political nature. We propose a method to derive several summary measures to tests for transformational accuracy when transforming input into output. Since dynamical flexibility of social algorithms prevents anticipating their behavior, we study these black boxes as if we study human behavior, through controlled experiments. We conceptualize them as noisy communication channels and evaluate their throughput with the same information theoretic measures engineers had originally used to minimize communicative distortion (i.e. mutual information). We use repeated experiments to reverse-engineer algorithmic behavior and test for its statistical significance. We apply the method to three artificial intelligence algorithms: a neural net from IBM's Watson, and to the recommender engines of YouTube and Twitter.

**Keywords:** algorithms, social media, recommender systems, information theory, mutual information, entropy, algorithmic behavior.

Growing at 25-30 % per year, the world's technological capacity to store and communicate information has grown too fast to be tamed by non- or even by semi-automated techniques (Hilbert, 2014, 2017, 2018). The silver lining is that the world's computational capacity has grown three times faster (with some 80 % per year, Hilbert & López, 2011). Humanity has taken advantage of this and long started to outsource the important task of interpreting and filtering digital content to computers with artificially intelligent algorithms.

Digital algorithms, defined as unambiguous digital recipes of how to transform input into output, have become superior to humans in the task of information mediation. For example, they have become better than human in image recognition (He, Zhang, Ren, & Sun, 2015), and speech recognition (Xiong et al., 2016), pushed by an word-error rate reduction from 26 to 4 percent just between 2012 and 2016 (Lee, 2016). Such interpretive power has given rise to omnipresent online recommender algorithms, which have become crucial gatekeepers in the management of today's communication landscape (Ricci, Rokach, Shapira, & Kantor, 2011). Their critical role has then again received much blame recently for creating filter bubbles and echo chambers that clearly restructure our communicational landscape (Bakshy, Messing, & Adamic, 2015; Colleoni, Rozza, & Arvidsson, 2014; Hilbert, Ahmed, Cho, Liu, & Luu, 2018; Pariser, 2011).

In this article, we use Shannon's "mathematical theory of communication" (1948) to derive summary measures that quantify different aspects of proactive algorithmic conversion between input and output of algorithms. This is important today, since traditional computer-mediated communication happened through rather transparent and passive channels. At the time when Shannon first conceptualized digital channels, the goal was to create noiseless channels, with as little distortion as possible, famously defined by Shannon's noisy-channel-coding theorem (Cover & Thomas, 2006). To the contrary, today's digital channels, especially in social media, are often highly proactive. Most of them are being actively shaped by algorithms that fundamentally pursue commercial interests (Lanier, 2018). This is sometimes more, sometimes less subtle.

Having a series of complementary summary measures that quantify how different the input is from the output is important in order to understand today's communication landscape. If input and output are identical, we have a passive and neutral channel, much in line with the channels of fixed line telephones from the time when Shannon developed his theory (1948). In this study, we use the same measures Shannon used to minimize noise in communication channels, but apply them to measure the nature and significance of existing transformations. For example, we feed YouTube with emotion-laden search terms, and evaluate how different is the emotional content of recommended videos. We follow users with certain personality profiles on Twitter, and study the arising patterns of personalities of recommended users to follow. In short, we model algorithms as noisy channels that intermediate between input and output and describe the statistical properties of the observed transformation with a long-standing summary numbers.

## Social Algorithms as Black Boxes

Many socially relevant algorithms have essentially become black boxes (O'Neil, 2017; Pasquale, 2015). Especially deep neural networks bury their functionality somewhere within up to hundreds of hidden layers (Castelvecchi, 2016; LeCun, Bengio, & Hinton, 2015). Additionally, the interplay of mutually influential social behavior and technological routines result in so-called social algorithms, which are too complex for anybody to understand their behavior fully, including those who programed them. The code might be deterministic, but social algorithms adjust their chosen behavior in real time to human dynamics, which makes their behavior as unpredictable the social phenomena it draws from. As such, "…the interplay of social algorithms and behaviors yields patterns that are fundamentally emergent. These patterns cannot be gleaned from reading code" (Lazer, 2015, p. 1090). Their creator may exert some unpredictable functionality intentionally in order to flexibly adapt to an ever-changing environment, while other aspects might be incidental and unintended.

This is unsettling, not only for social scientists who would like to understand today's computer-mediated communication, but also for private sector companies trying to improve their business models (Bakshy et al., 2015), policy makers trying to shape social development (Tutt, 2016; White House, 2016), and engineers trying to close back doors that give access to the possibility of manipulating their systems (Papernot et al., 2016).

The practical way forward consists in treating social algorithms as autonomous behavioral entities, which implies that we study them as we study human behavior. The aim is to reverse engineer their functionality as they function in their *natural habitat*, in order to better understand their *modus operandi* (Diakopoulos, 2015). First progress is being made in this regard. For example, Hannak et al. (2013) used repeated tests to piece apart the behavior of Google's search algorithm, and Hannak et al. (2014) use over 300 real-world accounts to understand price discrimination on 16 popular e-commerce sites. Mukherjee et al. (2013, p. 409) studied "What Yelp Fake Review Filter Might Be Doing?" and Guha et al. (2010, p. 81) analyzed the "Challenges in measuring online advertising systems". Also companies like Facebook themselves regularly undertake massive experimental efforts to understand the emergent behavior of the very own algorithms they us in public (Bakshy et al., 2015; Kramer, Guillory, & Hancock, 2014). While this is promising, a group of researchers recently cautioned about what they termed the "AI Knowledge Gap: the number of unique AI systems grows faster than the number of studies that characterize these systems' behavior" (Epstein et al., 2018, p. 1).

In this article, we offer one method to contribute narrowing this gap, contributing to the collective efforts that responds to the rising call that "Machine Behavior Needs to Be an Academic Discipline" (Iyad & Cebrian, 2018, p. 1). We study the black box of intelligent social algorithms with a rather simple, but time-honored method that quantifies throughput of noisy communication channels, i.e. how well and in what way the input matches the output. Given the

many unknowns of the modern algorithmic systems, we propose to use an approach that is well understood: information theory (Shannon, 1948). The contribution of this method consists in:

(a) capturing nonlinearities, which is crucial, given the unknown nature of the algorithmic transformation;

(b) calculating meaningful and complementary summary measures to elucidate different aspects of the transformation, and to compare behavior among different algorithms;

(c) testing for statistical significance against the null hypothesis that the detected throughput was the result of pure chance; and

(d) deducing additional conclusions from first principles (aka, 'theoretical deduction'), drawing from hundreds of related theorems and proofs from information theory (for a general overview of information theory see: Gleick, 2011; Pierce, 1980; for a more rigorous treatment: Cover & Thomas, 2006; MacKay, 2003).

To showcase the method, we apply it to three different artificial intelligence algorithms: a deep learning neural net of IBM Watson's natural language processing suite, and the socially embedded recommender systems from YouTube and Twitter.

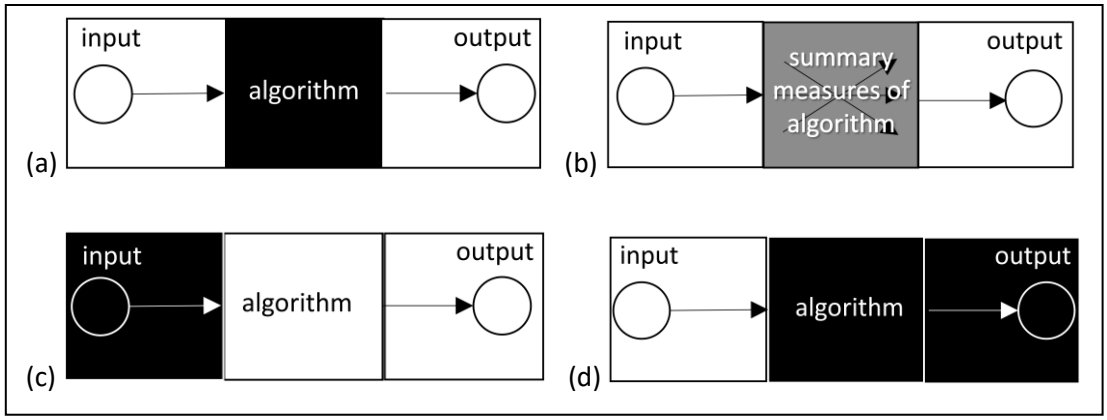
## Reverse Engineering of Algorithms

Any attempt to understand the behavior of an automaton implies reverse engineering its behavior. Reverse engineering is "the process of extracting the knowledge or design blueprints from anything man-made" (Eilam, 2011, p. 3). In software design, it means different things to different people (Chikofsky & Cross, 1990). It can refer to inferring the outline of code in detail (Richa, 2014), it can refer to a high level abstraction of the purpose of the software, satisfying managerial desire for control, while "having nothing to do with the actual design or construction of software" (Ensmenger, 2016, p. 323), and anywhere in between (Hall, 1992). Engineers who aim at improving the functionality of an algorithm will probably require more technical details (Papernot et al., 2016; Zahavy, Zrihem, & Mannor, 2016), while scientists might prefer less descriptive and more mathematical approximations of behavioral tendencies of algorithms (Bény, 2013; Mehta & Schwab, 2014). Doing social science, we are rather interested in options that are more abstract, and less of a technical replication. We take advantage of the fact that all algorithms must always have an input and output, which is a generalizable entry point for shedding light on the algorithm's black box behavior (Diakopoulos, 2015).

Different algorithms hide different aspects of their input-output relationship. Figure 1a is the prototypical black box, where we do have the ability to fully observe all inputs and outputs, but do not know what the algorithms does. Studying its behavior comes down to a 'Skinner box' like analysis of cause and effect with varying conditions. In some cases this input is tacitly collected, as is YouTube's watch-based recommendation engine (Davidson et al., 2010), while in other cases it is proactively solicited, as through likes (Castelluccio, 2006). In such recommender engines, the output is then again observable.

For reasons of completeness, it is important to mention that this is not the only black box scenario when studying algorithms (see Figures 1c and 1d). However, in our study, we focus on a first attempt to greying the black-box of Figure 1a, shown in Figure 1b. By collecting empirical data about input and output through controlled experiments, we shed lights on the throughput of the algorithm, which shows how well and in what way the algorithm's input matches its output. We do so by recording and calculating the involved joint, conditional, and marginal probabilities, which then allow us to calculate information theoretic channel properties (Cover & Thomas, 2006; MacKay, 2003; Shannon, 1948). This aims at providing a high-level summary assessment of the algorithmic transformation, more in line with behavioral science that does not aim at reverse engineering the neurological processes that govern behavior. However, the obtained summary measures are often all we need to know to direct social goals, especially when it comes to measuring the level of transformation in a channel. In terms of an analogy between complex algorithms and complex drugs, a health regulatory authority does not need to know the specifics of the drug receipts in order to determine if its outcome and effects are harmful to society (Tutt, 2016). Publishing the details would destroy the business model of the designer, while no assessment at all is social irresponsible. As such, this article contributes to the search for useful metrics that allow analyzing the overall effects of complex social algorithms.

Figure 1: Schematic scenarios of observability of algorithmic black-boxes. For Figure 1a and 1b see main text. In Figure 1c, we do not have access to the input. For example, "Amazon's recommendation secret" (Mangalindan, 2012, l. 1) hides the input, as it drives up to a third of the company's sales. While Amazon has recently even opened up the artificial intelligence behind its product recommendations (Klint, 2016), it has not opened up the 25 years of input data used to train its algorithms, which is the algorithm's main black-box. Figure 1d represents the case where only the input is transparent. For example, we might see our own mobility patterns on our Google timeline. We also know that many socio-demographic variables can be derived from trace data (Frias-Martinez & Virseda, 2013; Song, Qu, Blumm, & Barabási, 2010), but it is not clear what exactly companies do with this data input.



## Method

### Data and Procedure

We study three algorithmic communication channels and use our methods to compare their throughput behavior. In Case 1, we used emotion-laden key words as input and evaluated how deep learning NLP (natural language processing) interprets the meaning of these search terms. In Case 2, we fed a social video site (YouTube) with emotion-laden search terms and evaluated the emotional content of the transcripts of videos suggested by the online recommendation algorithm. In Case 3, we biased social media accounts (Twitter) toward certain personality traits, and then evaluated the personalities of the profiles recommended to follow on Twitter.

**Case 1: language processing channel.** Our first case is the natural language processing system AlchemyLanguage from the IBM Watson Developer Cloud (now also called Watson Natural Language Understanding). AlchemyAPI is an Application Programming Interface (API), originally launched in 2009 as a deep learning platform. Originally, it analyzed text as input for trading algorithms. In 2013, the company's software platform processed 3 billion API calls per month across 36 countries and in eight different languages (A. Williams, 2013). In 2015 it was acquired by IBM (IBM News, 2015). We focused on its identification of the so-called big five emotions in written text: anger, fear, disgust, joy, and sadness (Ekman, Sorenson, & Friesen, 1969; Philippot, 1993). As input, we looked for 100 synonyms of each of these five central terms in an online dictionary called 'reversedictionary.org'. We then fed the synonyms in random order into the artificial intelligence. As output, the tool assigns values between 0 and 1 to the presence of anger, fear, disgust, joy, sadness. We then normalized the sum of all scores to the total emotional charge.

Our hypothesis was that the output of this algorithmic transformation matches the input quite well. We expect there to be little noise in the transformation of words like 'funny' into the emotion of 'joy' and 'yikes' into the emotional category of 'fear'.

**Case 2: emotions recommender channel.** In the words of Google engineers, "YouTube represents one of the largest scale and most sophisticated industrial recommendation systems in existence" (Covington, Adams, & Sargin, 2016, p. 191). Therefore, in our second case, we biased the history of virgin YouTube accounts, using the same emption-based test terms as input. We then analyzed the emotional content of the transcripts of the recommended videos as our output.

In practice, we wrote a Python script that logged into a new account and searched for the first term from the list of 100 synonyms of one of the big five emotions. The script then watched the first seconds of the first recommended video, which prompts the system to move this videos

into YouTube's personal watch history.[1] After doing this for each of the 100 terms of a specific emotion, we collected the first 30 recommended videos.[2] Our script scraped the video title, description and the transcript of the words included in the video (the transcript was available for one third of the videos), and evaluated its emotional content, again with AlchemyLanguage.[3]

We hypothesized that the algorithm will somewhat relate input and output emotions. The social algorithm could perfectly match angry input into angry recommended videos, but it might also recommend videos with random emotional content, like joy and sadness. In each case, we would not know about the nature of the input-output conversion of emotional content. However, without a formal analysis, we were not sure if input and output relate to each other in a way that is significantly different from random transformations in a statistical sense.

**Case 3: networked personalities channel.** In our third case we biased virgin Twitter accounts by following users with an extreme personality type, and then evaluated the personality traits of the users suggested in the "Who to Follow" recommendations. For personality detection, we use IBM Watson Personality Insights service, which was trained to detect five prominent personality traits of Twitter profiles. We focus on its Big Five personality traits (also known with the acronym OCEAN, or NEO- with some add-on (Costa & McCrae, 1976)), which is the most widely used personality model (Costa & MacCrae, 1992; McCrae & Costa, 1987). It evaluates a person's degrees of openness (experience variety), conscientiousness (organization and thoughtfulness), extraversion (stimulation seeking), agreeableness (compassion and cooperation), neuroticism (emotional range and sensitivity).[4]

We randomly sampled 1,500 user ID numbers from Twitter's API first one million ID numbers, and worked with the 1,484 user profiles for which the Personality Insight suite provided a valid results (some profiles had insufficient content (less than the 1,500 words required by IBM's solution), or are in unsupported languages). We created five virgin Twitter

---

[1] It justifies to only watch the first recommended video as it has been shown that the highest ranked search results are exponentially more likely to be clicked than lower ranked links (Bakshy, Messing, & Adamic, 2015). It is also important to start watching the video, as we found that YouTube's recommendation algorithm works on basis of the watch history, not on basis of the search history. We speculate that the reason is that the final consumption of online content is a mix of own search results and of input from their online friends (Gottfried & Shearer, 2016; Hilbert, Ahmed, Cho, Liu, & Luu, 2018).

[2] We decided to stick to this rather smaller sample size, because users focus on the top recommendations (Bakshy et al., 2015; Gottfried & Shearer, 2016), and because we found that videos further down the list are only loosely related to the search content (at the time of the study in 2017, after about 180 recommended videos, suggestions started to repeat).

[3] After some preliminary testing, we decided to evaluate the feelings separately for each title, description and the transcript (if available), and then to build the simple average of these three groups. This reduces biases due to the length of the text (i.e. gives more weight to titles, which are otherwise overwhelmed by the transcript).

[4] IBM's Personality Insights service infers personality characteristics by representing words as vectors in a highly dimensional space through an unsupervised learning algorithm called GloVe, which is part of the so-called Word2Vec family (Pennington, Socher, & Manning, 2014). It works with the ratio of the probabilities of co-occurring words and outperforms other machine learning and dictionary based models.

accounts and had each one following the 20 highest scoring users in each of the big five personality characteristics, respectively. These five stylized Twitter accounts were our input. We then obtained the top 15 recommended user profiles of the "Who to Follow" suggestion list from Twitter's recommender engine and evaluated their personality traits.[5]

Our hypothesis was that the algorithm skews the recommendations toward the stylized personality of the users already followed by the biased account, but not by much. Personality traits are only one out of many possible variables to consider, but being on broadcast media like Twitter, it seems like personality traits of those to follow could be an interesting factor to consider. Nevertheless, yet again, without a formal analysis, we were not sure if input and output relate to each other in a significantly significant way.

**Measures**

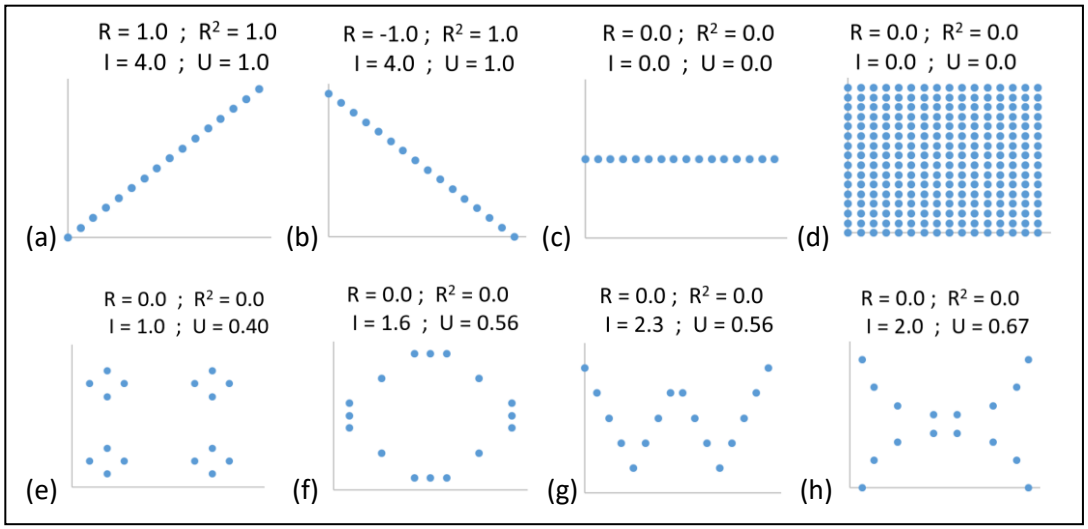
We use the traditional information theoretic channel setup and interpret the input as sender and the output as receiver (for a formal introduction to communication channels see Cover & Thomas, 2006, p. Ch. 7). In his noisy-channel-coding theorem, Shannon (1948) derived the measure of mutual information to establish an upper limit of non-distorted communication over a noise contaminated channel. The legendary theorem says that it is possible to communicate error-free over a noisy channel up to the maximum of the channel's mutual information.[6] The mutual information between the random variables of a sender input $S$ and a receiving output $O$ is a symmetrical measure of association denoted with $I(S;O)$ (Cover & Thomas, 2006).

From a measurement perspective, using information theory for our purposes has two main benefits. The first is that information theoretic measures, like mutual information, naturally capture nonlinearities. This is important, since it would be very limiting to assume from the onset that complex social algorithms exclusively perform linear transformations. Figure 2 compares the conventional linear measure of Pearson's correlation coefficient R with mutual information *I*, and with its normalized version, which we call *U*. We can take the horizontal x-axis as the input and the vertical y-axis as the output of the transformation. For example, the case in Figure 2a implies a noiseless and non-distorted transformation between sender and receiver. The graph shows that both measures capture equally well associations that are either strong or non-existent (Figures 2a–d). However, as shown in Figures 2e–h, the linear measure fails to capture non-linear transformations, while the information theoretic measure captures them.

[5] We average the normalized personality scores reported by IBM's Personality Insight among our users (20 input and 15 output users), and then normalize among the achieved percentiles of the five traits (between 0 and 1).

[6] Interestingly, what is known as the 'Shannon limit' was not achieved until the so-called turbo-codes of the mid-1990s (Berrou, Glavieux, & Thitimajshima, 1993), almost half a century after Shannon showed that is must exist.

Figure 2: Comparison of Person correlation coefficient with mutual information, given input and output variables with 16 possible realizations.




Much like a covariance, which is the basic ingredient for Pearson's correlation coefficient, mutual information measures the difference between the joint and the independent distribution, but uses the logarithm (and therefore their ratio) of the involved probabilities of the random variable of a sender input $S$ and a receiving output $O$ (see equation (1).[7] It is measured in bits when the base of the logarithm is 2

$$I(S;O) = \sum_{s,o} p(s,o) * \log_2 \left( \frac{p(s,o)}{p(s) * p(o)} \right) \qquad (1)$$

An alternative way to calculate mutual information shows the second main benefit, namely that it is part of a group of several meaningful measures that are all complementary to each other and highlight different aspect of a communication channel. These are different entropy measures. If mutual information is akin to the covariance, entropy is akin to the variance of a variable. Entropy reaches its maximum value with a uniform distribution and its minimum (zero) when all probability density is placed on one single realization of the random variable. Therefore, entropy measures the level of uncertainty or uniformity of the probability distribution (equation 2).

$$H(S) = -\sum_{s} p(s) * \log\big(p(s)\big) \qquad (2)$$

The mutual information can be understood as the intersection between two entropies. It measures how much information one variable contains about the other, and vice versa. This is often visually represented as the overlapping intersection in the form of the Venn diagram shown

[7] As customary, we use capital letters to refer to random variables, like $S$ and $O$, and its minuscular counterparts to refer to concrete realizations of that variable, like $s$ and $o$.

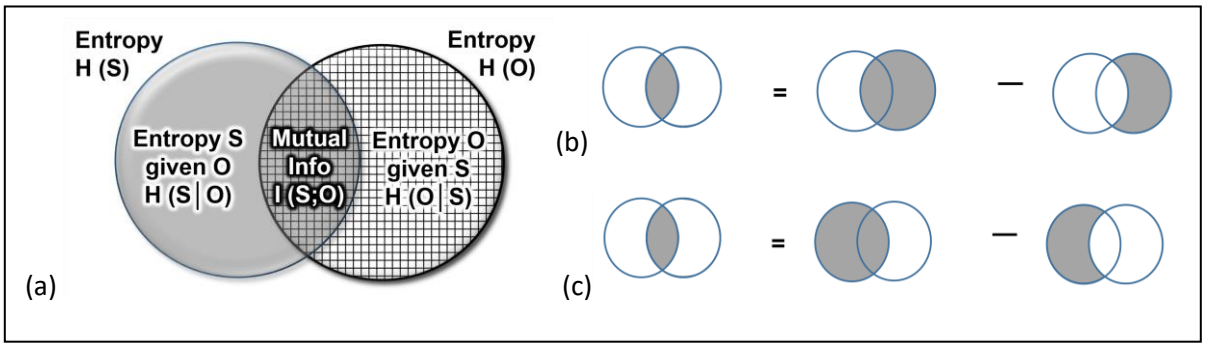
in Figure 3a, also called i-diagrams (James, Ellison, & Crutchfield, 2011; Yeung, 1991). Breaking the mutual information down into its entropy components allows to quantify the level of noise from input to output (measured by the conditional entropy $H(O|S)$ (equations 3a–b)), as well as the level of equivocation, from output to input (measured by the conditional entropy $H(S|O)$ (equation 3c)) (Pierce, 1980). The noise of a channel is the distortion of the channel when viewed from the perspective of the input: given the input, how different is the output? The equivocation looks at the channel the other way around, from the perspective of the output, and asks: given the output, how different is the input? Both conditional measures are fundamentally related by Bayes' theorem, whose far-reaching history underlines the notorious trap to erroneously equate both perspectives. They are complementary to each other and emphasize different aspects of the same process.

$$H(S|O) = -\sum_{s,o} p(s,o) * \log\big(p(s|o)\big) \qquad (3a)$$

$$I(S;O) = H(O) - H(O|S) \qquad (3b)$$

$$I(S;O) = H(S) - H(S|O) \qquad (3c)$$

Figure 3: Schematic illustration of information theoretic metrics in the form of information diagrams. Circles represent entropies, intersections the mutual information. Figures 3b and 3c match equations 3b and 3c, respectively.




Note that mutual information, like entropy, is not a normalized measure. It depends on how many different realization or categories of the variable. The uncertainty of the outcome of the roll of a dice with six possibilities is simply larger than the uncertainty inherent to a binary coin flip with only two choices. In order to make it comparable among channels with different numbers of input and output categories, one can normalize the mutual information with the entropy of the input: $U = \frac{I(S;O)}{H(S)}$. This is sometimes called the uncertainty coefficient, coefficient

of constraint, proficiency, or entropy coefficient (Press, Teukolsky, Vetterling, & Flannery, 2007, p. 761), and asks: what fraction of the input *S*, is preserved in the output *O*?[8]

Summing up, information theoretic measures are summary measures that allow us to conveniently express the average level of noise or distortion for all incoming- and/or outgoing-variables. For the purpose of communication, one can equate the concepts of noise and distortion. Note that these measures represent averages over all realizations and do not automatically quantify which of the realization has which effect on the overall distortion, which is often an important question asked in studies interested in biases or social discrimination.

### Statistical Tests

Following the common rigor of social science research, we need to make sure that our results are not mere artifacts of random chance. Since no parametric distribution of errors is known for the nonlinear measure of mutual information, we need suitable surrogate data to test the null hypothesis of independence between the input and output. There are several ways to do this.[9] We create randomized control groups as input, and compare their throughput to our result in a one-sided significance test. If our empirically detected mutual information is frequently larger than in the random-input channel, we can say with confidence that it is unlikely that the obtained throughput is the result of pure chance. If it is not, then the null hypothesis is not rejected and the output of the channel is statistically independent from the input. We might as well have fed the channel with some random input, and get the same type of output. We randomize only the input to preserve trivial dependencies of the channel. This aims at destroying the dependency of the outgoing transition probability on the input.

[8] Note that there are several ways to define the uncertainty coefficient. One could also normalize on the output *O*, which measures the fraction of the output, or on the joint entropy of both variables. These are not necessarily the same, and explain different fractions. Only normalization on the smaller of both entropies can reach a coefficient normalized between 0 and 1 (since mutual information is their intersection, see Figure 3a).

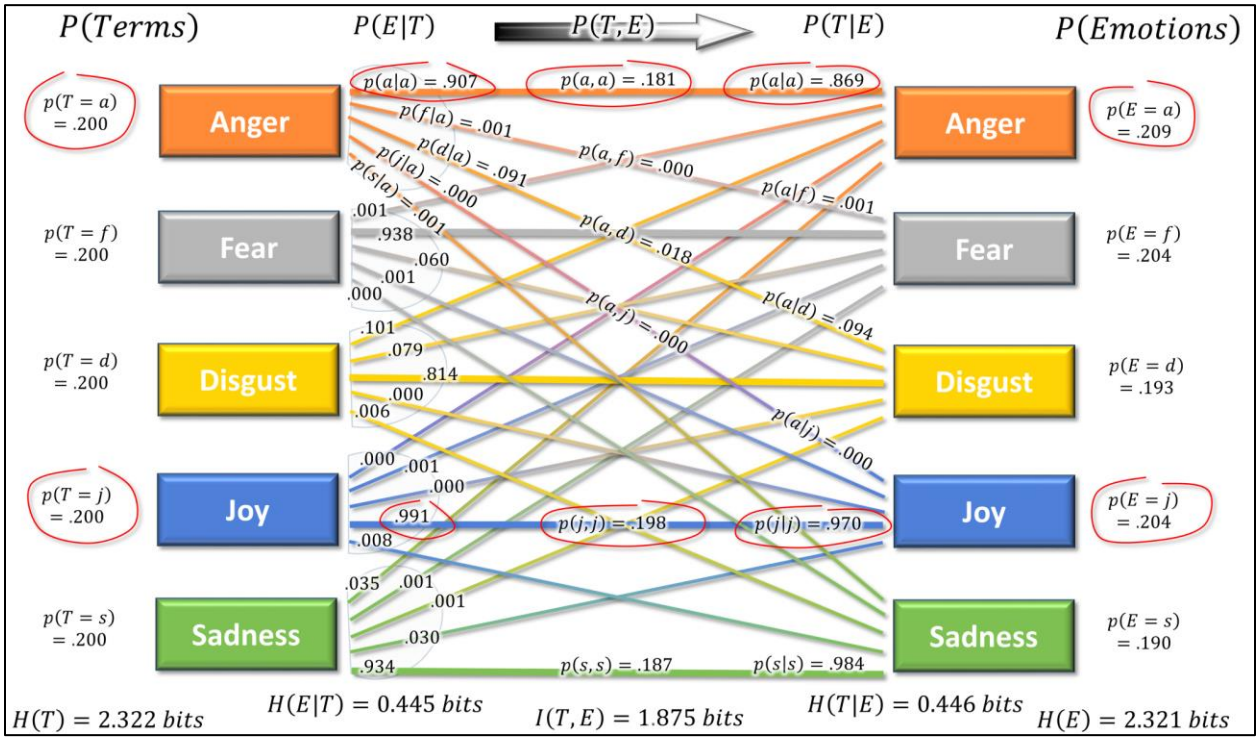
[9] If we would have pair-wise observations of input and output as joint events (such as [sad user & sad user], [sad user & angry user], etc.) we could simply bootstrap different aspects of our channel by permutation randomization with the goal of destroying dependency among the variables (Chávez, Martinerie, & Le Van Quyen, 2003; Han, 1980; Hilbert, Ahmed, et al., 2018). In our case, however, we do not have input-output pairs, but rather tendencies of normalized scores (such as [user x% sad & user y% sad]). Therefore, here we suggest checking the significance of our information flows with something more akin to a randomized control group. The bad news is that, in contrast to simple bootstrapping, control groups require more work-intensive experiments. The good news is that they provide a more comprehensive picture the of channels algorithmic behavior.

## Results

### Case 1: Language Processing Channel

Feeding 100 synonyms of one of the big five emotions into AlchemyLanguage creates a conditional random variable $P(E|T)$ that asks: given search term $t$, what distribution of emotions $E$ is perceived by the artificial intelligence?[7] As expected, we found that the algorithm's output identified the equivalent emotion with high probability. For example, we found $p(e = anger|t = anger) = 0.907$, $p(fear|a) = 0.001$, $p(disgust|a) = 0.091$, $p(joy|a) = 0.000$, and $p(sadness|a) = 0.001$ (see Figure 4).

Figure 4: Characterization of the emotion-based input and output of the AlchemyLanguage natural language processing algorithm as a communication channel.

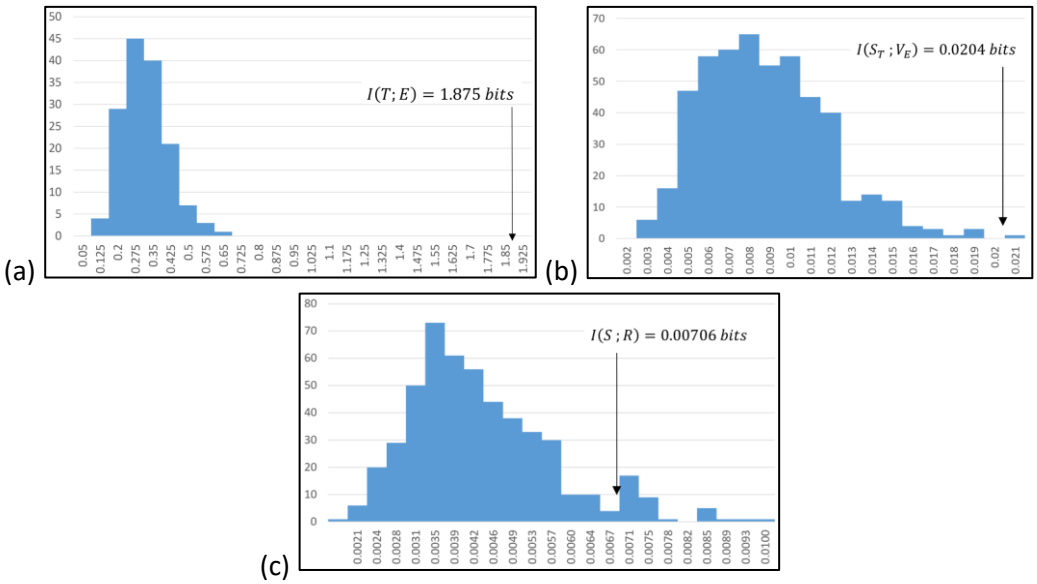



We used the same number of synonym terms for each of the big five emotions. This means that we chose a uniformly distributed input variable, $P(T)$. In information theory, (diagonal) crossover transitions from input to output are understood as noise, and (horizontal) throughput transitions as so-called identity-, or mutual information transitions (see Figure 4). If the channel would be noiseless, the horizontal identity transitions would carry 100% of the transition probability, and all diagonal crossover transition would have zero probability. In our case, the identity transition is the largest transition, transmitting between 81.4 % and 99.1 % of the input correctly (e.g. see circled case for anger and joy in Figure 4). The existing noise distorts the output distribution. In general, this is shown by the fact that the output distribution is less

uniform (and therefore with less entropy) than the input distribution: $H(T) = 2.322\ bits > H(E) = 2.321\ bits$. In specific, the emotions of anger, fear and joy became overrepresented, while disgust and sadness are underrepresented (Figure 4). This shows that this algorithmic channel slightly distorts the input-output flow of emotions.

Since we know both the probability of the input and, from our experiment, the transition probability, so we can also calculate the joint probability, $p(t) * p(e|t) = p(t, e)$, which makes it straightforward to calculate the mutual information with the help of either equation (1) or equation (3b): $I(T; E) = 1.875\ bits$. We can also use Bayes' theorem to calculate the 'equivocation', the probability of having received some input, given a certain output: $P(T|E) = P(E|T) * \frac{P(T)}{P(E)}$. It turns out that from the perspective of the output, we have slightly more uncertainty about the input than the other way around: $H(E|T) = 0.445\ bits < H(T|E) = 0.446\ bits$. The finding that noise is smaller than the equivocation makes sense for our algorithmic setup, since it means that the channel is less uncertain when going 'from input to output' than vice versa (but it is by no means necessary or automatic). The mutual information is more than four times larger than both the noise and the equivocation in the channel: $I(E; T) \gg H(E|T) \approx H(T|E)$. There is much more accurate throughput, than distortion. The normalized uncertainty coefficient, $\frac{I(T;E)}{H(T)} \approx \frac{1.875}{2.322} \approx 0.807$, tells us that the algorithmic transformation maintains more than 80% of the input. The channel still distorts the throughput, but the output also has a clear informational relationship with the input.

Figure 5: Feeding randomized control groups into the algorithmic channel. Distribution of mutual information for: (a) 150 randomized input trails for the case 1 Language Processing Channel; (b) 500 randomized input trails for case 2 Emotions Recommender Channel; (b) 500 randomized input trails for case 3 Networked Personality Channel.



For our bootstrapped significance test, we randomly pick 100 search terms from the total collection of 500 synonyms (containing 100 synonyms of each of the five emotions). We repeat the experiment with such randomized input 150 times. Figure 5a shows that the mutual information of the resulting 150 channels is much lower, implying that the channels are much noisier than our original channel, which had much stronger identity transitions. It is therefore very unlikely that our obtained throughput is part of the family of randomized control channels. Given this clear result, we stopped running control experiments after 150 repetitions, so being exact, we can say that there is at least a chance of $p = \frac{1}{150} = 0.0067 < 0.01$ that the empirically obtained mutual information is larger than with random channel input.

**Case 2: Emotions Recommender Channel**

We then did a similar experiment with YouTube's video recommender system. We fed it with the same terms as in the previous case, now called $(S_T)$, consisting of the 100 synonyms fir an emotion, and obtained the emotional content from the transcripts of the recommended videos (as described above), resulting in the random variable $(V_E)$ (see Figure 6). Naturally, we expected the algorithmic transformation in this channel to be noisier than in our previous case, since we now do not evaluate the emotions of the search terms directly, but the emotions of the text associated with videos that result from those search terms.

Figure 6 shows the resulting overall channel. We calculate the same information theoretic metrics as before. We obtain a much lower mutual information between input and output $I(S_T, V_E) = 0.0204\ bits$. Our uncertainty coefficient barely: $\frac{I(S_T,V_E)}{H(S_T)} \approx 0.009$. There is clearly more noise than in the previous case, with $H(V_E|S_T) = 2.14\ bits$. Taking a closer look at the data reveals that for most cases, the noiseless horizontal identity transition is not even the most likely outgoing transformation. Regardless of the input, the most likely transformation results in "Joy". For example, feeding 100 synonyms of "anger" into the channel, the output shows videos that are associated 42 % with joy (only 18% with anger), while feeding in 100 synonyms of "sadness" results in 37 % joy-related video output (only 24% with sadness) (see underlined marks in Figure 6). This leads to the fact that the majority of the emotions associated with the recommended videos relate to joy, 38.7 % of them.

We returned to the raw dataset to look for possible explanations, and suspect that this can be partially explained by social influence. We carried out the relevant data collection between March 29 and April 1, 2017. Since April 1 is "April fools' day" in many Western cultures, including the U.S., where we ran this study, many of the recommended videos contained content related to pranks, practical jokes, and hoaxes. It is easy to imagine that on this day input terms from our anger list, such as "annoyed", "mad", and "provoke", can result in videos with rather joyful content full of jokes aimed at provoking laughter. This suggests that these algorithms are social algorithms, whose *modus operandi* is only interpretable in light of the surrounding social

influence. The behavioral outcome might make sense in hindsight, but is difficult to predict without systematic behavioral experiments focused on the intertwined social behavior of the mediated system.

Figure 6: Characterization of the emotion-based input and output of YouTube's recommender system algorithm as a communication channel.

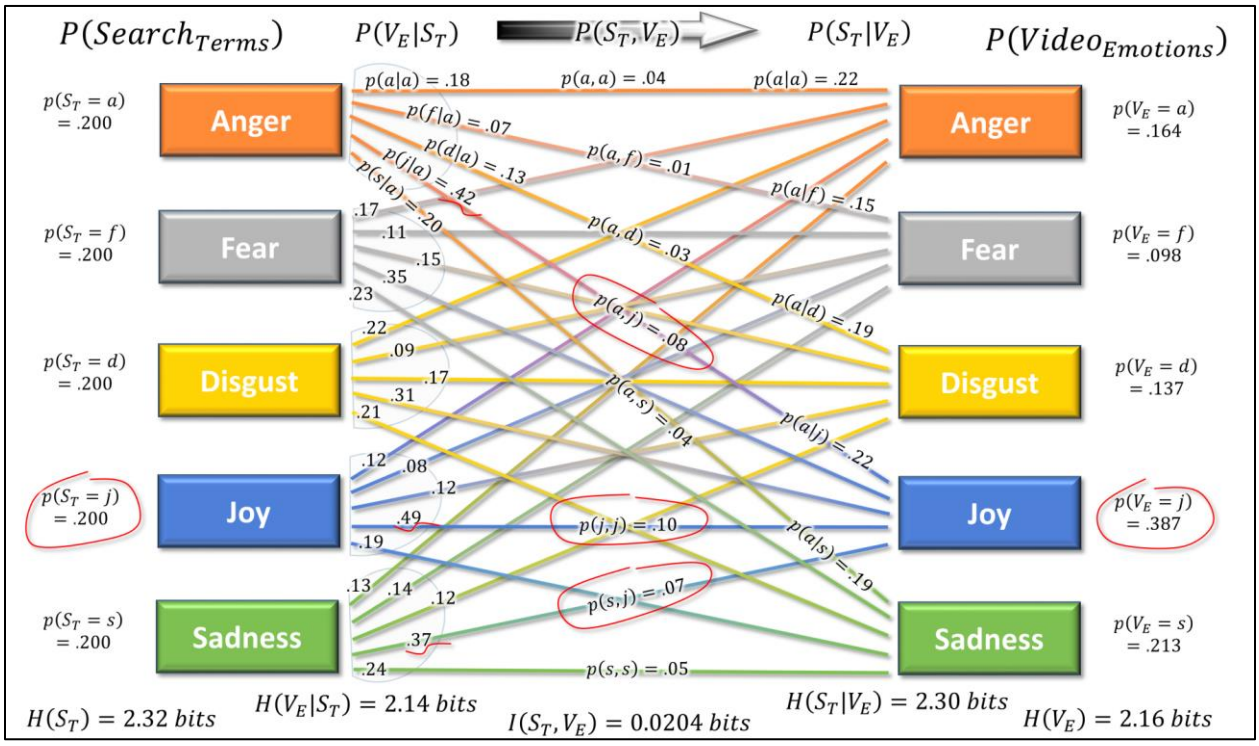



Figure 5b shows the distribution of mutual information for 500 randomized trial experiments for this case. We find one randomized control group that has a mutual information that is higher than the one we empirically detected: $I(random; R) = 0.0205 > 0.0204 = I(T; R)$. This means that we found a chance of 1 in 500 (or $p = 0.002$) that our channel obtained the detected channel throughput through pure luck. Following common convention, we can say that also in this case the detected information is statistically significant, as it is larger than what is randomly expected at the level $p < 0.01$, but not at $p < 0.001$.

**Case 3: Networked Personalities Channel**

As a third case, we analyzed how the follower recommender system of Twitter processes personality traits. At the outset, we did not know if personality plays a role in the matchmaking of people on this micro-blogging service, but it seemed to be intuitive that it could be one of the candidate traits to select for when recommending whose tweets to follow.

We use IBM Watson's normalized personality scores to create a communication channel between personality traits of the users that we followed in five stylized accounts (our source $S$), and the personality traits of recommended users to follow (the recommendation $R$). Table 1 presents the source distribution and the outgoing noise transition probabilities, as measured by our experiment. All other probabilities can be calculated from these measured frequencies with the laws of probability, including $P(S,R)$, $P(S|R)$, and $P(R)$ (i.e. it is the equivalent of the previous graphs in table format).

Table 1: Noisy transitions of networked personality channel: source $P(S)$ and transition $P(R|S)$.

| s | p(s) | | p(r \| s) | | | | | p(R\|s) |
|---|---|---|---|---|---|---|---|---|
| ↓ input | output → | | Agr. | Consc. | Extr. | Neuro. | Open. | |
| Agreeableness | 0.194 | | 0.15 | 0.20 | 0.22 | 0.19 | 0.24 | 1.00 |
| Conscientiousness | 0.197 | | 0.14 | 0.23 | 0.18 | 0.19 | 0.27 | 1.00 |
| Extraversion | 0.200 | | 0.15 | 0.17 | 0.23 | 0.19 | 0.26 | 1.00 |
| Neuroticism | 0.203 | | 0.11 | 0.20 | 0.22 | 0.23 | 0.25 | 1.00 |
| Openness | 0.205 | | 0.14 | 0.17 | 0.20 | 0.19 | 0.30 | 1.00 |
| | 1.000 | *p(r):* | *0.14* | *0.19* | *0.21* | *0.20* | *0.26* | |

Our input distribution $P(S)$ is not uniform, which is not necessary. Information theory allows to feed any kind of distribution into our channel. In this case, our 20 users with the most extreme personality scores in agreeableness achieve on average a lower score than the 20 users with the most extreme personality scores of openness. Our Twitter users seem to be slightly more open than agreeable.

Conditioned on the five stylized accounts with specific personality traits, the recommendation algorithm distributes follower suggestions in a quite similar fashion. Conditioned on the input, agreeableness is the least probable personality trait (with $P(r = agr. \mid S)$ between 11% and 15%) and openness the most likely (average of 26%). It makes intuitive sense that a social media algorithm suggests following people with an open personality. At the same time, the algorithm also recommends to follow more neurotic users than agreeable user, which is interesting to note and invites to speculations. For all traits, exactly 19% of the recommendation scores are aimed at neuroticism (lower and slower emotional range), except for the identify transition from neuroticism to neuroticism, which is a bit higher. The identity transition (found on the diagonal of the transition matrix of Table 1) is the highest or second highest conditional probability for each personality trait, except for agreeableness, where it is the lowest.

The entropies of the input and output are more similar than in the case of the previously analyzed YouTube recommender system: $H(S) = 2.32\ bits$ and $H(R) = 2.29\ bits$. However, just because both are similarly uniform they are not necessarily also related. In fact, the mutual

information is very low, with $I(S;R) = 0.00706\ bits$, which gives us our lowest uncertainty coefficient: $\frac{I(S;R)}{H(S)} \approx 0.003$. This originates from the uniformity of the distribution within the channel. If the identity transition would be dominant (the diagonal entries in Table 1), input and output would be more strongly related.

Given such low level of throughput, we will certainly have to check if it is just the result of mere chance. Figure 5c shows the distribution of mutual information for 500 randomized trials for this case. Aiming for a comparable bootstrap, we kept the slightly skewed input distribution $P(S)$, and tested what kind of profiles would result if we drew 20 profiles randomly from the same pool of input users that we used in the original experiment. This aims at destroying the causal dependency of throughput, while preserving trivial dependencies. As shown in Figure 5c, we find that exactly 35 out of the 500 random draws have higher mutual information, achieving a communication transmission as high as 0.01 bits, compared to the 0.007 bits of our experiment. This is 7 % of our surrogate distribution, $p = \frac{35}{500} = 0.07 > 0.05$. Traditional statistical conventions would argue that we cannot reject the null-hypothesis that our experiment is just a case of random chance. It is not statistically different from sending random input through the channel. While the exact cut-off of significance levels are certainly subject to debate (some scholars might argue for $p > 0.1$), it is certainly striking that several of our truly random channel obtained higher information throughput that our original channel, which we manipulated with highly specialized content in terms of personality profiles. This shows, at the very least, that matching personality profiles are not a top priority for Twitter's recommender algorithm.

## Discussion

### Comparing different channels

Keeping things simple, we worked with variables with the same number of input and output categories (five each). This allows us to compare meaningfully the absolute measures of entropies and mutual information between different cases directly. Otherwise, we would have to rely on the normalized uncertainty coefficient alone for comparative purposes.[8] We obtained quite different values for the informational input-output conversions in our channels:

- Case 1 (natural language processing channel): $I(T;E) = 1.875\ bits$ ($p < 0.01$), with a normalized uncertainty coefficient of some 81 %;
- Case 2 (emotions recommender channel): $I(S_T, V_E) = 0.0204\ bits$ ($p = 0.002 < 0.01$), with an uncertainty coefficient of some 0.9 %.
- Case 3 (networked personalities channel), $I(S;R) = 0.007\ bits$ ($p = 0.07 > 0.01$), with an uncertainty coefficient of some 0.3 %.

Our first algorithmic channel clearly has very little active conversion among the identified variables. This was to be expected, since this neural net was explicitly trained for the

variable-matching we tested for in our experiment. Comparing cases 2 and 3 underlines the old wisdom that in statistical tests it is important to consider both significance and effect size. Our bootstrapped significance tests show that the transformation from our input into our output, is statistically different from random input for case 2, but not for case 3. At the same time, in both cases the transformation preserves less than 1% of the information contained in the input (measured in bits). This is quite a high level of distortion. This being said, our method allows us to quantify both the level of statistical significance and the effect size and give the researcher the opportunity to draw conclusions.

This is useful, not only and necessarily as an ultimate goal, but as also input for further explorations. First, today's digital communication channels are rarely clean and passive channels, especially when going through social platforms. Previous generations of scholars did not have to worry about proactivity of channels, and most literature in computer-mediated-communication still does not pay enough attention to the active role of algorithms. Knowing that and if, then how different the input is from the output, summarized in a simple number, is valuable in its own right.

Secondly, most Communication scholars are not ultimately interested in understanding social algorithms per se. They are more interested in exploring the effects of social algorithms in different communication process. For example, they are interested if a non-biased search input leads to illegal discrimination in the output (Caliskan, Bryson, & Narayanan, 2017; Hajian, Bonchi, & Castillo, 2016), if different variables play an active role in the algorithmic personalization or not (Hannak et al., 2013, 2014), or how much skewed political opinions get even more or less skewed after filter-bubble recommender engines played their omnipresent role in polarization (Bakshy et al., 2015; Colleoni et al., 2014). Our method does not provide all details of such transformations, but it allows scholars to quickly and broadly identify how well the input matches the output, which is an important first assessment that was not necessary only a few years ago, when communication channels were still passive and clean. This assessment can then be used to inform posterior explorations of variables that measure social effects.

**Insights from information theoretic theorems**

One last benefit of using information theoretical measures is that it allows us to stand on the shoulders of giants, those that built the theoretical fundament of digital communication on basis of hundreds of theorems and proofs within this framework (Cover & Thomas, 2006; MacKay, 2003). This allow us to deduce additional conclusions from first principles (aka, 'theoretical deduction').

**The limits of predictability.** For example, we can derive the limits of predictability from entropy measures (Hilbert, James, Gil-Lopez, Jiang, & Zhou, 2018; Song et al., 2010). Information theory, which can be seen as the theoretical fundament of probability theory, and vice versa (Jaynes, 2003), allows us to do this independent from the specificities of the predictive method, be it traditional extrapolation, the most recent cutting-edge neural network, or the yet

undiscovered next big thing in artificial intelligence. They are and will all be subject to Fano's inequality, which relates the probability of error in guessing the random variable *O* to its conditional entropy $H(O|S)$ (see equations 3a-3c) (Cover & Thomas, 2006, Chapter 2.10). No predictive algorithm can be better than the limit of predictability $\Pi$:

$$\Pi = 1 - \frac{H(O|S) - 1}{log_2|A|} \qquad (4)$$

Where $|A|$ is the number of categories of the output (the number of different choices we have in our prediction), in our case five. Plugging numbers of our three cases into equation (4) shows that we will not be able to predict the outcome of the algorithmic transformation of case 2 with more than 51 % accuracy and of case 3 with 45 % accuracy. Fano's inequality does not provide any reason why it should not be possible to predict case 1 with 100 % accuracy.[10] Being a lower bound, this does not mean that we must hit these limits of predictability necessarily and certainly not automatically. However, it tells us that we certainly cannot do better. The channel simply does not carry more information than what is bound by Fano's inequality.

**Theoretical multivariate decomposition.** In practice, when analyzing complex social algorithms, we often deal with multivariate communication channels. Social algorithms often subsume many variables, some of them are left out or even have to be left out, since they are even hidden from external evaluation. Information theoretic theorems can be useful in the approximation of some of the properties of such these hidden variables, even without the required empirical data.

For example, the algorithm of the emotions recommender channel of case 2 does not recommend emotions, but videos. More relevant to our specific analysis, it recommends different text transcript of videos. We then derive emotions from transcripts of recommended videos. In our previous analysis, any bias or distortion stemming from the transcript (which is often automatically produced) was not accounted for. Let us call unevaluated variable of the video text $V_T$. This leaves us a third stylized variable, two are measured directly ($S_T$: search terms; and $V_E$: video emotions), and (at least) one that is unevaluated. In this case, this variable is not hidden, in a sense that we have it, but we have not undertaken any effort to quantify it in any other way than with its emotional content, so there could be all kinds of confounding variables and biases connected to it.

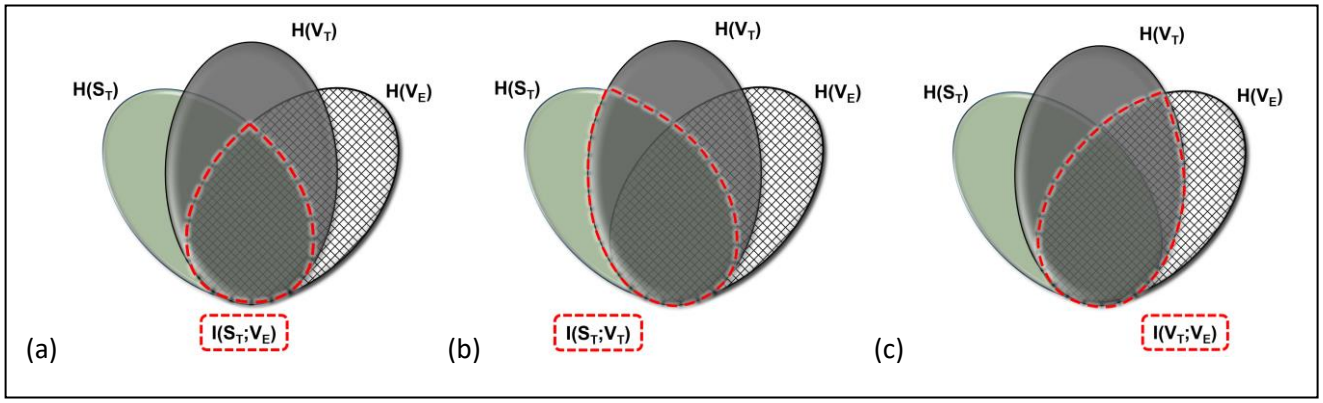
In theory, a three variable setup leads to 7 multivariate relations $(2^3 - 1)$, which can be depicted with a multivariate information diagram (see James et al., 2011). In our particular case, we can assume that we have a special constellation of this setup. We cannot go from the search terms to the resulting video emotions without passing through the video text (see Figure 7). As a result, the variable of the video text $V_T$, shields the input search terms $S_T$, from the output video

[10] Note that Fano's inequality gives a lower bound on the limit of predictability, and does not give any indication how right this bound is. Actually, equation (4) results in 1.24 for Case 1, which would imply that one could make predictions with 124% accuracy, which is of course nonsensical.

emotions $V_E$. In theory, the original and final variables have nothing in common that was not mediated through the intermediating variable $V_T$. In other words, search terms $S_T$ and video emotions $V_E$ are conditionally independent, conditioned on video text $V_T$. This implies that there is no mutual information between the input and output that is not mediated by the shielding variable: $I(S_T; V_E|V_T) = 0$.[11] This kind of Markovian shielding can often be expected when modelling algorithms as channels: some key variable completely intermediates in the transformation from input to output. Note that in practice, this conditional independence might not exist. In this case, the algorithm might use something else, unrelated to the transcribed text, which relates input and output.

One mathematical theorem that takes advantage of conditional independence is known as the data-processing inequality (Cover & Thomas, 2006). It states that the existence of mediating variables can never result in information gain, and will usually result in a loss of information from input to output. Intuitively, the data-processing inequality says that no clever manipulation of information can increase the amount of information that is being processed. Processing of information cannot get more out of it than was originally in it: informational output can never exceed informational input. This a useful formal result when trying to understand the properties of the algorithmic black boxes we are dealing with.

Figure 7: Theoretical multivariate decomposition of three variables involved in case 2, showing that the outlined intersection in (a) $I(S_T; V_E)$, must be smaller than the ones outlined in (b) $I(S_T; V_T)$, and in (c) $I(V_T; V_E)$.




More formally, it says that if the random variable $S_T$ communicates with the mediating variable $V_T$, which then communicates with the final variable $V_E$, (in other words: if $S_T \rightarrow V_T \rightarrow V_E$ is conditionally independent), then the mutual information between the original and the final variables $I(S_T; V_E)$ can never be larger than the mutual information between adjacent variables: $I(S_T; V_E) \leq I(S_T; V_T)$ and $I(S_T; V_E) \leq I(V_T; V_E)$. In our case, we then already know that both

[11] This effectively reduces the number of multivariate relations from 7 to 6.

$I(S_T; V_T)$ and $I(V_T; V_E)$ must be larger than $0.0204\ bits$, even so we never observed $V_T$.[12] The visualizations in Figures 7a – c can be used as a guide to prove the data processing inequality, a proof that is visually quite intuitive.

## Conclusion: benefits and drawbacks

As a contribution to the larger goal of developing formal method to quantify the behavior of social algorithms, we presented a method to quantify the conversion between the input and output of algorithmic transformations with a few simple summary variables. Knowing if and how much the input of a social algorithm differs from the output is important nowadays, because digital channels are rarely passive and neutral, but rather proactive mediators. We showed that working with information theoretic measures has four main benefits, namely (a) capturing nonlinearities, which means minimal methodological assumptions about the nature of the ongoing transformation; (b) calculating meaningful and complementary summary measures, such as the mutual information, the level of noise and equivocation; (c) the possibility to test for statistical significance additionally to the involved effect size; and (d) the possibility to deduce additional conclusions from first principles, based on the far-reaching theory that underlies digital communication, aka the "mathematical theory of communication" (Shannon, 1948).

Being a methodological exploration, the specific details of our case studies were not as important to us as the demonstration of the behavior of the measures. For example, we created our main variables with help of automated semantic analysis, which is never exact. While comparative tests against human evaluations of both emotions and personality place both of the tools we used on the cutting edge semantic analysis from textual data (IBM Bluemix, 2017; Hilbert et al., 2017), the question what exactly is measured is for our purposes less important as the coherence between the evaluations at both ends of the channel.

The same accounts in general for the variables that we chose, which, in future studies will have to receive much attention. Naturally, the definition of the identified variables and the chosen method of their measurement frames the result. As with traditional behavioral experiments, it is up to the researcher to identify the variables that matter. This does not change. We could have formulated different variables to test for different aspects of distortion, and measured our variables differently. For example, there are different ways to classify emotions (Holyst, 2017), including a recent suggestion that people evaluate 27 distinct emotions when consuming online videos (Cowen & Keltner, 2017). Additionally, it is likely that the designers of YourTube's recommender engine might not even have considered emotions explicitly. It was certainly not their goal is to match emotional content, but rather to engage users, which can be done by considering a battery of the most diverse variables. From our reserve-engineering

[12] This relation holds both ways, which can be formalized with stating that $S_T \leftrightarrow V_T \leftrightarrow V_E$ forms a Markov Chain (they relate to each other in a consecutive order under conditional independence). The equation shows equality if the mediation is noiseless.

perspective, there are countless aspects one could test for when trying to understand what social algorithms do what they do. The application of our methods is mute on the choice of input and output variables. Still, our method might help to discard candidate variables, as it can be applied systematically to test for the significance and effect size of candidate variables on distortion. Even though YouTube's recommendation might not be have been designed from the outset to process emotions, our tests show that it does. This is a useful finding, be it the result of intentional design, or an emergent externality. We showed that it does have a measurable effect.

As for more fundamental limitations, it is important to note that this approach faces the same challenges of stationarity as all other experimental assessments of dynamic phenomena. Modern social algorithms are dynamic and neither match input against a static library, nor do they rely on a fixed environment. This leads to different results when assessing their behavior at different times. The behavior of algorithms varies in their degree of stationarity, and therefore, in the generalizability of the results obtained with our analysis. By the time this study is published, the YouTube recommender engine might already work completely different. We noted that even the behavior of the natural language neural net from IBM Watson changed somewhat, not within the a few days, but within a time window of several weeks. Today's social algorithms even adjust language interpretation dynamically while the social environment changes. More rigorous tests of algorithmic stationarity will be required in future research, and information theory might again be able to help in that challenge (Kennel & Mees, 2000).

Finally, while there are benefits to using the proposed summary measures of entropy and mutual information, there are also well-known drawbacks. One, notoriously already mentioned by Shannon upfront, is that the most straightforward application of information theory works for categorical variables, whereas each category does not automatically carry any meaning (see also McKinney & Yoos, 2010). In contrast, variance and correlations from traditional statistics work for scalar variables, and we at least know that a 3 is larger than a 1. This limitation can be seen when comparing Figures 2a and 2b: mutual information does not distinguish between a positive or negative correlation. It is more in line with $R^2$, than with R (see Figures 2a – b). There are extensions of information theory to ordinal differences and scalar variables, but they are not as straightforward and there is currently no consensus on the best way to go about it (Corominas-Murtra, Fortuny, & Solé, 2014; Crutchfield, 1991; Plotkin & Nowak, 2000). Something similar applies to multivariate analysis. While linear statistics has successfully be scaled to an arbitrary number of variables (Monge & Cappella, 1980), information theorists are still search for scalable solutions (James, Emenheiser, & Crutchfield, 2019; P. L. Williams & Beer, 2010). This shows that more methodological work will be needed.

In this sense, the presented method will certainly not solve all challenges involved in quantifying the ever more omnipresent role of intelligent algorithms, but it aims at intensifying an impending discussion about different ways to greying the opaque behavior of black-box algorithms. The specific contribution aims at the quantification of the algorithmic distortion between two variables.

## References


Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Bény, C. (2013). Deep learning and the renormalization group. *ArXiv:1301.3124 [Quant-Ph]*. Retrieved from http://arxiv.org/abs/1301.3124

Berrou, C., Glavieux, A., & Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Technical Program, Conference Record, IEEE International Conference on Communications, 1993. ICC '93 Geneva* (Vol. 2, pp. 1064–1070 vol.2). https://doi.org/10.1109/ICC.1993.397441

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Castelluccio, M. (2006). The Music Genome Project. *Strategic Finance*, *88*(6), 57.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20. https://doi.org/10.1038/538020a

Chávez, M., Martinerie, J., & Le Van Quyen, M. (2003). Statistical assessment of nonlinear causality: application to epileptic EEG signals. *Journal of Neuroscience Methods*, *124*(2), 113–128.

Chikofsky, E. J., & Cross, J. H. (1990). Reverse engineering and design recovery: a taxonomy. *IEEE Software*, *7*(1), 13–17. https://doi.org/10.1109/52.43044

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Corominas-Murtra, B., Fortuny, J., & Solé, R. V. (2014). Towards a mathematical theory of meaningful communication. *Scientific Reports*, *4*. https://doi.org/10.1038/srep04587

Costa, P. T., & MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Incorporated.

Costa, P. T., & McCrae, R. R. (1976). Age Differences in Personality Structure: a Cluster Analytic Approach. *Journal of Gerontology*, *31*(5), 564–570. https://doi.org/10.1093/geronj/31.5.564

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd Edition). Hoboken, NJ: Wiley-Interscience.

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198). New York, NY, USA: ACM. https://doi.org/10.1145/2959100.2959190

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Crutchfield, J. P. (1991). *Semantics and Thermodynamics* (Santa Fe Institute Working Paper). Retrieved from http://www.santafe.edu/research/working-papers/abstract/ebcd34a934704fc1af231314b3af1432/

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., … Sampath, D. (2010). The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 293–296). New York, NY, USA: ACM. https://doi.org/10.1145/1864708.1864770

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Eilam, E. (2011). *Reversing: Secrets of Reverse Engineering*. John Wiley & Sons.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science (New York, N.Y.)*, *164*(3875), 86–88.

Ensmenger, N. (2016). The Multiple Meanings of a Flowchart. *Information & Culture: A Journal of History*, *51*(3), 321–351. https://doi.org/10.1353/lac.2016.0013

Epstein, Z., Payne, B. H., Shen, J. H., Dubey, A., Felbo, B., Groh, M., … Rahwan, I. (2018). Closing the AI Knowledge Gap. *ArXiv:1803.07233 [Cs]*. Retrieved from http://arxiv.org/abs/1803.07233

Frias-Martinez, V., & Virseda, J. (2013). Cell Phone Analytics: Scaling Human Behavior Studies into the Millions. *Information Technologies & International Development*, *9*(2), 35–50.

Gleick, J. (2011). *The Information: A History, a Theory, a Flood*. New York: Pantheon.

Gottfried, J., & Shearer, E. (2016, May 26). News Use Across Social Media Platforms 2016. Retrieved December 23, 2016, from http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

Guha, S., Cheng, B., & Francis, P. (2010). Challenges in Measuring Online Advertising Systems. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (pp. 81–87). New York, NY, USA: ACM. https://doi.org/10.1145/1879141.1879152

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2125–2126). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2945386

Hall, P. A. V. (1992). *Software Reuse and Reverse Engineering in Practice*. London, UK, UK: Chapman & Hall, Ltd.

Han, T. S. (1980). Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, *46*(1), 26–45. https://doi.org/10.1016/S0019-9958(80)90478-7

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring Personalization of Web Search. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 527–538). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2488388.2488435

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 14th ACM/USENIX Internet Measurement Conference (IMC'14)*. Vancouver, Canada. Retrieved from http://personalization.ccs.neu.edu/PriceDiscrimination/Research/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification (pp. 1026–1034). Presented at the Proceedings of the

2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society. https://doi.org/10.1109/ICCV.2015.123

Hilbert, M. (2014). How much of the global information and communication explosion is driven by more, and how much by better technology? *Journal of the Association for Information Science and Technology*, *65*(4), 856–861. https://doi.org/10.1002/asi.23031

Hilbert, M. (2017). Information Quantity. In *Encyclopedia of Big Data* (pp. 1–4). Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_511-1

Hilbert, M. (2018). Communication Quantity. In *Encyclopedia of Big Data* (pp. 1–6). Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_512-1

Hilbert, M., Ahmed, S., Cho, J., Liu, B., & Luu, J. (2018). Communicating with Algorithms: A Transfer Entropy Analysis of Emotions-based Escapes from Online Echo Chambers. *Communication Methods and Measures*, *0*(0), 1–16. https://doi.org/10.1080/19312458.2018.1479843

Hilbert, M., James, R. G., Gil-Lopez, T., Jiang, K., & Zhou, Y. (2018). The Complementary Importance of Static Structure and Temporal Dynamics in Teamwork Communication. *Human Communication Research*. https://doi.org/10.1093/hcr/hqy008

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*(6025), 60–65. https://doi.org/10.1126/science.1200970

Holyst, J. A. (Ed.). (2017). *Cyberemotions - Collective Emotions in Cyberspace*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43639-5

IBM Bluemix. (2017). *The science behind the service* (Bluemix Docs, Personality insights). Retrieved from https://console.bluemix.net/docs/services/personality-insights/science.html

IBM News. (2015, March 4). IBM Acquires AlchemyAPI, Enhancing Watson's Deep Learning Capabilities [CTB10]. Retrieved April 10, 2017, from https://www-03.ibm.com/press/us/en/pressrelease/46205.wss

Iyad, R., & Cebrian. (2018, March 29). Machine Behavior Needs to Be an Academic Discipline. Retrieved October 3, 2018, from http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline

James, R. G., Ellison, C. J., & Crutchfield, J. P. (2011). Anatomy of a bit: Information in a time series observation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*(3), 037109. https://doi.org/10.1063/1.3637494

James, R. G., Emenheiser, J., & Crutchfield, J. P. (2019). Unique Information and Secret Key Agreement. *Entropy*, *21*(1), 12. https://doi.org/10.3390/e21010012

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. (G. L. Bretthorst, Ed.) (1 edition). Cambridge, UK ; New York, NY: Cambridge University Press.

Kennel, M. B., & Mees, A. I. (2000). Testing for general dynamical stationarity with a symbolic data compression technique. *Physical Review E*, *61*(3), 2563–2568. https://doi.org/10.1103/PhysRevE.61.2563

Klint, F. (2016, May 16). Amazon's Giving Away the AI Behind Its Product Recommendations. *WIRED*. Retrieved from https://www.wired.com/2016/05/amazons-giving-away-ai-behind-product-recommendations/

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Lanier, J. (2018). *Ten Arguments for Deleting Your Social Media Accounts Right Now*. Henry Holt and Company.

Lazer, D. (2015). The rise of the social algorithm. *Science*, *348*(6239), 1090–1091. https://doi.org/10.1126/science.aab1422

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, A. (2016, March 18). Geoffrey Hinton, the "godfather" of deep learning, on AlphaGo. *Macleans.Ca*. Retrieved from http://www.macleans.ca/society/science/the-meaning-of-alphago-the-ai-program-that-beat-a-go-champ/

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms* (1 edition). Cambridge, UK ; New York: Cambridge University Press.

Mangalindan, J. P. (2012, July 30). Amazon's recommendation secret. Retrieved April 7, 2017, from http://fortune.com/2012/07/30/amazons-recommendation-secret/

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90.

McKinney, E. H., & Yoos, C. J. (2010). Information About Information: A Taxonomy of Views. *MIS Quarterly*, *34*(2), 329–344.

Mehta, P., & Schwab, D. J. (2014). An exact mapping between the Variational Renormalization Group and Deep Learning. *ArXiv:1410.3831 [Cond-Mat, Stat]*. Retrieved from http://arxiv.org/abs/1410.3831

Monge, P. R., & Cappella, J. N. (1980). *Multivariate techniques in human communication research*. New York : Academic Press. Retrieved from http://trove.nla.gov.au/version/45226159

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What Yelp Fake Review Filter Might Be Doing? In *Seventh International AAAI Conference on Weblogs and Social Media* (pp. 409–418). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. S.l.: Penguin.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical Black-Box Attacks against Machine Learning. *ArXiv:1602.02697 [Cs]*. Retrieved from http://arxiv.org/abs/1602.02697

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin.

Pasquale, P. of L. U. of M. F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Philippot, P. (1993). Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition & Emotion*, *7*(2), 171–193. https://doi.org/10.1080/02699939308409183

Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise* (2nd Revised ed. (1st 1961)). New York, NY: Dover Publications.

Plotkin, J., & Nowak, M. (2000). Language Evolution and Information Theory. *Journal of Theoretical Biology*, *205*(1), 147–159. https://doi.org/10.1006/jtbi.2000.2053

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender Systems Handbook*. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-85820-3

Richa. (2014, June 6). Reverse Engineering Tutorial: How to Reverse Engineer Any Software. Retrieved April 7, 2017, from https://blog.udemy.com/reverse-engineering-tutorial/

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423, 623–656. https://doi.org/10.1145/584091.584093

Song, C., Qu, Z., Blumm, N., & Barabasi, A.-L. (2010). Limits of Predictability in Human Mobility. *Science*, *327*(5968), 1018–1021. https://doi.org/10.1126/science.1177170

Tutt, A. (2016). *An FDA for Algorithms* (SSRN Scholarly Paper No. ID 2747994). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2747994

White House. (2016). *Preparing for the Future of Artificial Intelligence*. Washington D.C.: Executive Office of the President National Science and Technology Council Committee on Technology. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Williams, A. (2013). AlchemyAPI Raises $2 Million For Neural Net Analysis Tech, On Par With IBM Watson, Google. Retrieved April 10, 2017, from http://social.techcrunch.com/2013/02/07/alchemy-api-raises-2-million-for-neural-net-analysis-tech-on-par-with-ibm-watson-google/

Williams, P. L., & Beer, R. D. (2010). Nonnegative Decomposition of Multivariate Information. *ArXiv:1004.2515 [Math-Ph, Physics:Physics, q-Bio]*. Retrieved from http://arxiv.org/abs/1004.2515

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., … Zweig, G. (2016). Achieving Human Parity in Conversational Speech Recognition. *ArXiv:1610.05256 [Cs]*. Retrieved from http://arxiv.org/abs/1610.05256

Yeung, R. W. (1991). A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, *37*(3), 466–474. https://doi.org/10.1109/18.79902

Zahavy, T., Zrihem, N. B., & Mannor, S. (2016). Graying the black box: Understanding DQNs. *ArXiv:1602.02658 [Cs]*. Retrieved from http://arxiv.org/abs/1602.02658