Reducing Subjectivity and Bias in an Officer's Analysis of Suspicion in Drug Interdiction Stops

Arthur Crivella
Founder, Crivella Technologies, Ltd
Pittsburgh, Pennsylvania USA
acrivella@crivellatech.com

Wesley M. Oliver[1]
Professor of Law
Duquesne University School of Law
Pittsburgh, Pennsylvania USA
oliverw@duq.edu

Morgan A. Gray
Third Year Law Student
Duquesne University School of Law
Pittsburgh, Pennsylvania USA
graym2@duq.edu

ABSTRACT[2]

Police officers must daily determine whether they have justification to hold cars they have stopped for ordinary traffic investigations for further investigation. Yet these determinations involve the interpretation of very fact-specific case law that do not yield predictions for subsequent cases and are fraught with subjectivity if not actual bias. Artificially intelligent systems hold the potential to lessen the impact of implicit biases by assisting officers in making these decisions with greater consistency on the basis of factors relevant to suspicion. Using patented text recognition algorithms in order to identify content of interest, or relevant language, our prototype is capable of reading case law and police reports to identify factors relevant to suspicion. With this information, the likelihood a court approve a search or detention can be assessed. Police reports identifying the bases for fruitful and unsuccessful searches will then permit the system to assess the odds that drugs are present. Finally, by identifying race-neutral language uniquely used to describe suspicious circumstances involving minority motorists, the effect of implicit biases in the officer's description can be mitigated.

Keywords. machine learning, artificial intelligence, unstructured data, prediction, implicit racism, racial bias, end user vernacular, vector regression, content of interest, text recognition, saliency, probable cause, reasonable suspicion

1. Introduction

Fact-specific standards for assessing suspicion give officers little guidance. Perhaps worse, the malleability of the standards allows discriminatory decisions to hide. Even when decision-makers do not expressly consider race in their decisions, implicit biases affect outcomes when legal rules lack the clarity to constrain discretion. Machine learning, however, offers the prospect to consistently apply the standard, even the possibility of identifying and ignoring language that implicitly identifies race in a suspicion analysis. Drug interdiction stops provide an excellent corpus of decisions to begin to train computers to evaluate suspicion as they involve a relatively small number of variables. Early testing reveals that such a system could be deployed to assist officers in deciding whether summoning drug dogs in traffic stops is appropriate. Preliminary analysis of the corpus of decisions further suggests that at least some implicit officer biases find their way into written descriptions of the facts. The algorithm will be taught to identify race-unique language and discount any extent to which it contributes to a finding of suspicion.

2. Unpredictable and Unreliable Legal Standards for Assessing Suspicion

Probable cause permits an officer to search a car for drugs; on reasonable suspicion he can detain it until a drug dog can sniffed around it for the presence of drugs. These tests do not themselves give officers guidance. An officer is said to have probable cause justifying a search if the situation would "warrant a belief by a [person] of ordinary caution that a crime has been committed." [1] Reasonable suspicion is a belief, based on "specific and articulable facts," as well as "rational inferences from those fact" that a crime has been committed, or is being committed. [2] The ambiguity in these definitions is obvious on their face. One unfamiliar with search and seizure law would not even be able to identify which of the two standards is more difficult to satisfy, much less apply these standard to the circumstances of a stop.

Certainly no multi-factor test provides predictability or certainty. Decisions of courts offer guidance only in so far as the facts in those cases are analogous. Occasionally the facts of a case will be practically indistinguishable from a subsequent case. More often, however, an officer, or attorney advising the officer be unable to locate a predictably analogous case in a short order. Computers, unlike humans, are capable not only of searching and identifying cases with the similar factors, they are capable of identifying the

[1] Corresponding Author: Wesley M. Oliver, Duquesne Law School, 900 Locust Street, Pittsburgh, PA 15282, E-mail: oliverw@duq.edu

weight to be assigned factors identified by officers in assessing suspicion. They are capable of evaluating the sufficiency of suspicion with a level of understanding no human could.

There is another problem with these legal standards that machine learning can ameliorate. There is no way for officers, or judges, to meaningfully predict the likelihood that drugs are present on the basis of a given set of facts. Probable cause and reasonable suspicion are standards relating to the odds that a crime has been committed. Yet courts have rejected any efforts to put numbers to suspicion. Understandably, courts have shied away from efforts to express these standards in terms of numbers that cannot be meaningfully generated. As in other areas of technology, however, such as DNA matching, where scientific advances have enabled odds calculations, the law quickly accepted the use of statistical probabilities. [3] An empirical method of assessing suspicion is similarly within the grasp of modern technology and would provide a more rational way of justifying a search than our current system that necessarily relies on the codified hunches of judges.

An algorithm for assessing suspicion has the potential to reduce the biases that have frequently been alleged to be a part of drug interdiction efforts. Obviously by producing results that better predict the presence of drugs, fewer fruitless searches will occur of all racial groups, lessening the false-predictive effect of biases. In the early work with the corpus of judicial decisions, however, we have been able to identify language that, while neutral on its face, is strongly correlated to descriptions of motorists or passengers of minority racial groups. The computer can be programmed to ignore suspicious inferences that it would otherwise identify from this language officers tend to use more often for particular groups of motorists.

3. An Ideal Legal Test for Machine Learning and Prototype Implementation

A relatively small number of factors inform an officer's assessment of whether drugs are likely to be a car that has been stopped for an ordinary traffic offense. An algorithm for evaluating suspicion is therefore not overly complex. Additionally data sets that include successful and fruitless searches already exist that describe the officer's basis of suspicion. Some data therefore exists enabling an assessment of the odds that drugs are present.

Police departments across the country engage in drug interdiction on the highway. The patterns of drug interdiction on the highway is always the same. An officer stops a motorist for some traffic offense and then determines whether these is a basis to hold the car long enough for a drug dog to sniff the car – or search the car for drugs even in the absence of a positive canine hit. Hundreds of thousands of judicial opinions exist, in federal and state courts, that resolve the question of whether an officer has observed sufficient circumstances to justify either continued detention or a search.

Additionally, the New Jersey State Police are in possession of data that includes officers' basis for conducting searches in drug interdiction stops that led to both searches yielding drugs as well as searches that were unsuccessful. As a result of claims of racial profiling on the New Jersey Turnpike, the state police entered into a consent decree with the United States Department of Justice which required them to record the reason for every stop, use of a drug dog, and search of an automobile. [4] Officers were further required to record whether contraband was discovered in any searches conducted. This training data will provide the system the ability to begin to assess the question of the odds that drugs are present – a separate question from whether a court would find a search or detention is justified.

Only a very small number of factors can be part of an officer's assessment of suspicion this context. Of course officers can see or smell an illegal drug. For non-trivial suspicion evaluations, the short list of potential bases of suspicion will include: (1) the smell of potential masking agents (talcum powder, cologne, air fresheners, or baby wipes, for instance); (2) travel plans that are inconsistent between passenger and driver or inconsistent with a rental agreement; (3) travel plans to or from cities known as source or wholesale locations for drugs; (4) car makes and models identified as drug courier cars; (5) nervous behavior or furtive gestures by the driver or passenger; and (6) luggage in the back seat, potentially to make room for drugs in the trunk. This small number of factors makes training relatively easy.

Implementation of this system can be effected with little difficulty and will allow refinement of the system. Typically an officer assessing suspicion has an opportunity to return to his patrol car during the stop while his dispatcher determines whether the motorist has any outstanding warrants. At this point the officer would have an opportunity to describe the circumstances he identifies that are of concern. As it presently exists, officers using this system can learn the likelihood that courts would find a search or detention justified and will soon be able to learn the probability of the presence of drugs based on the data from the New Jersey Turnpike. In the pilot programs, however, the system will become ever more-sophisticated.

The criteria of suspicion in drug interdiction stops are not binary, though many current police reports treat them as if they were. Prompts for officers allow the system to fine-tune what it has already learned about the significance of these criteria. Nervousness is not something that is either observed or not. A conclusion of nervousness is supported by observations of behavior. When officers indicate that a motorist or a suspect is nervous, the system will prompt the officer to describe the actions she observed that led her to that conclusion. Further, nervousness is not something that a suspect exhibits or lacks. When an officer identifies this as a factor supporting her suspicion, the system will ask the officer to grade the level of suspicion, much like a pain scale, on a scale of 1 to 10. Prompts are similarly appropriate for other factors such as odors and inconsistent travel stories. These prompts then allow the training of the system to evaluate the presence of suspicious factors in the degree to which they are identified, using the officer's rating scale (which can be weighed against the scores assigned to this factor in other cases by both the officer and the department), as well as the qualitative descriptions offered by the officer.

4. Developing a Prototype to Asses Suspicion[3]

In order to develop the suspicion analyzing prototype we have created two corpra of data. The first, we have named the Judicial Probable Cause Opinion Corpus. To develop it, we first began by using existing case law consisting comprised of language from judicial opinions from state and federal appellate and trial courts. Each case was read and annotated using *Knowledge Kiosk*, a system developed by Crivella Technologies Limited. To begin, over 100 judicial opinions were uploaded to *Knowledge Kiosk* for annotation. These opinions were hand selected and carefully scrutinized to ensure relevance. The opinions focus on drug interdiction stops, and whether or not officers had the requisite reasonable articulable suspicion that criminal activity was afoot to detain the motorist for a longer period that was necessary to effect the stop.

After reading these cases several factors upon which officers commonly rely in making their determinations were identified, including but not limited to; nervousness, masking agents, inconsistent stories, rental vehicles being used, etc. The pertinent texts were annotated into bigrams, trigrams, and larger word clusters. These word clusters have been collected and will become part of a separate algorithm using a semantic word similarity model based on latent semantic analysis to identify parallel language in unannotated texts. This semantic word similarity model has been described in detail below.

This first set of documents which we annotated, described above as the "Judicial Probable Cause Opinion Corpus," is what we have identified as a reference corpus. It is suitable because it contains a wide variety of language pertinent to the description language that identifies factors upon which offers rely to make a suspicion determination. From this reference corpus, we identify pertinent language, divide and excerpt it into bigrams, trigrams, and larger word clusters, which are then put together into a seed corpus. These word clusters, which we call seed markers, have been compiled to create a separate seed corpus upon which other corpora of language, or test corpus, will be tested for content of interest.

> was traveling 82 in a 75 mile per hour zone. During the stop Patrolman Rettinger became suspicious of Defendant because he appeared unusually nervous, his luggage was in the back seat instead of the trunk, and there was an unusual quantity of fast food wrappers on the passenger floorboard of the vehicle. Patrolman Rettinger also noticed discrepancies in Damato's answers to where he had rented the car and where he was headed to. Damato told Patrolman

Figure 1: Language of Officer Suspicion annotated from a judicial opinion. [5]

Figure 1 is an excerpt from a judicial opinion containing the raw language relevant to creating a seed corpus, The language identified here was broken up into seed marks and markers to create a seed corpus that was used for reference in determining content of interest from test copra.

The seed corpus comprises a plurality of textual units and each textual unit of the seed corpus comprises at least one instance of a seed marker included in the seed marker set. When constructing seed markers they may be based on previous statistical analysis of corpora, experience of the user, etc. Here, they have been derived from the language of judicial opinions. To ensure validity after the seed corpus was generated it was analyzed to verify that the seed markers within in face return content of interest text. At step statistical values for the seed markers in the seed corpus were calculated. Statistical values used to describe seed markers used were frequency, z-score, and saliency. For the statistical values that require comparison to a reference corpus we used a previously existing corpus of language that has been used in existing corpuses derived from a litigation scenario, and is statistically valid itself. The saliency of a marker may be found by multiplying the average of the marker's corpus and textual unit bases z-scores and rarities according to the following equation:

$$\text{Saliency} = \frac{freqz+filez}{2} * \frac{freqz+filez}{2}$$

Saliency is a statistical value that may be generated with respect to one marker or marker set, and may describe the significant of an occurrence of the marker or marker set based its frequency and the frequency of other similar markers in the subject corpus. Where freqz is the z-score of the marker by frequency in the corpus; filez is the z-score of the marker by the number of textual units including instances of marker; freqr is the rarity of the seed marker by frequency in the corpus; and filer is the rarity of the seed marker by number of textual units including instances of the marker.

Following the creation of the seed corpus, a separate evaluative algorithm was created. This algorithm was created in order to deploy to seek content of interest in unannotated documents.

The development team had to consider the size of the corpus and strategize on the best way to proceed with the data. It is possible that for large subject corpora, it may be impractical to compute some or all of the statistical values discussed above due to limitations relating to processing speed, memory requirements, etc. Accordingly, some or all of the statistical values above may be generated based on a stratified sample of a subject corpus. The sample may be a selection of textual units chosen from the subject corpus such that all of the textual units in any given subset of the subject corpus have an equal chance of selected for the sample, and

[3] All information relied upon in creating this system has been derived from the patented system developed by Crivella Technologies Ltd, U.S. Patent No. 7,779,007 B2.

such that no subsect of the subject corpus is disproportionately represented in the sample. In this way, the statistical properties of the sample may mirror those of the subject corpus as a whole. Sampling the subject courses may not be necessary for all types of statistical analyses, or in all cases. For example, if the desired statistical analysis is not processor and/or memory extensive, and/or if the corpus is relatively small, then sampling may not be required. It will also be appreciated that as process and memory technology improves, the need for sampling will obviously lessen. In our case, the subject corpora was not too large and improvements in technology have allowed us to analyze thousands of cases in an instance.

Following the creation of the seed corpus comprised of the set markers derived from the judicial opinions, an evaluative algorithm was created to determine content of interest in test corpora.

The various elements identified are directed to methods of identifying content of interest within an unannotated corpus of language. These methods comprise the step of applying a first marker set to the corpus, where the first marker set comprises at least one marker identifying a first type of text. More specifically, various elements have applied one or more markers or marker sets, to test a corpus to identify content of interest. However, in this instance, it is related to language of suspicion.

A set of evaluable rules was developed to describe the content of interest sought in unannotated texts, or test corpora. Evaluative rules are of various types. For example, evaluative rules are binary and/or quantitative. Binary rules are defined a threshold criterion and candidate textual unites either meet or fail to meet (e.g., certain metadata criteria, particular scores for a given number or marker set, etc.) Binary rules are expressed inclusively or exclusively. For example, under an inclusive expression, candidate textual units that meet a threshold may be considered likely to include content of interest. Under an exclusive expression, candidate textual units that meet a threshold are eliminated from further consideration. Quantitative rules rate the likelihood that a given textual unit contains content of interest based on a predefined criterion or set of criteria. For example, a textual unit having a score range of scores for a given marker set or sets may be considered to have a predetermined likelihood of including content of interest.

The evaluative rule set include one or more rules that consider the results of applying a marker set identifying text of a particular type to the test corpus. For example, the application of a marker set to the test corpus yields a raw score that indicates the number of occurrences of markers in the marker set in the corpus, and/or in each textual unit of the corpus. Where any markers in the marker set are weighted, the raw score is weighted accordingly, The raw score itself is a criterion of one or more of the evaluative rules (e.g., if the raw score for marker set A is less than X then eliminate it from consideration). Also, various values derived from the raw scores of the marker sets make up evaluative rule criteria. Exemplary derivative values include, a z-score for the marker set based on its application to a reference corpus, a rarity of the marker set, and saliency of the marker set.

Here, and as described above in more detail, our marker sets are the language of interdiction and officer vernacular. The method of identification we have chosen is as follows. Textual units within the testing corpus, in this case the annotated texts of the judicial opinions and officer language have been assigned a score for each marker set and that score that indicates a degree to which the textual unit includes text of the type identified by the marker set. Following this, a set of evaluative rules was applied to the unannotated textual units, based in part on their scores to determine their content of interest language. For example, different marker sets may include markers identifying text an event. However, in this case, they contain markers that identify language of suspicion. See Figure 2 below.

By applying similar methods we will create a second corpus of language which we have named the "Language of Active Interdiction." This will contain language identical and similar to the language that an officer of the law. Our approach has been to identify criminal complaints and other similar statements from officers describing the facts identified above as they seem them and in the common vernacular that will have described them as it may and will likely differ from a judicial opinion. We will use a similar method described above to identify content of interest language in a breadth of unannotated text.

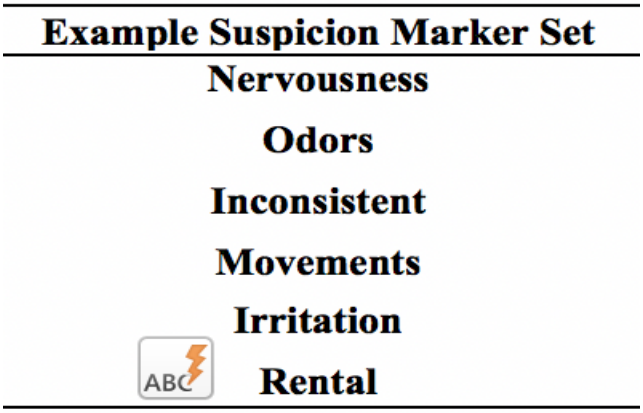

Figure 2: Example of a simple marker set containing language of suspicion.

5. The Prototype to Assess Suspicion

We have begun developing a prototype system. This system is still in development, however, it is nearly ready to deploy. We have first identified the categories of factors that officers commonly rely upon in make suspicion assessments. We have selected the categories as follows: Vehicle Status, Occupant Behavior, Vehicle Contents, Travel Route, and Prior Offenses as shown below in figure 3. In each of these broad categories more specific language is included. For example, under Vehicle Status, set markers like "rental" or "major drug highway" would be included.

Each of these categories has been broadly drawn to include many specific instances that would fit under each category. For example, under the category of occupant would fit, sweating, heavy breathing, shaky voice, avoiding eye contact, etc. Each separate scoring category will, based on the Judicial Probable Cause Opinion Corpus will be programmed to recognize each subcategory therein. Each subfactor will be assigned a specific value. This value will depend on the legal weight the factor is decided in making the legal analysis based on the jurisdiction of the officer. The ultimate decision on whether or not reasonable suspicion has been met to detain the motorist will ultimately be decided by whether or not the factors present have reached the threshold.

ID	Name	Vehicle Status	Occupant Behavior	Vehicle Contents	Travel Route	Prior Offenses	Score
669	DAMATO, Nicholas	1	2	2	1	no	1.5
670	MASON, Victor Eugene	3	5	3	2	yes	3.25
671	BULLOCK, Michael Antonio	1	1	0	0		0.5
672	PRADO, Angel	1	2	2	2		1.75
673	Taylor,	1	1	1	1		1
674	RIBBLE, Amethyst	2	4	0	0		1.5
675	WROBEL, Michael John	2	3	1	1		1.75
676	Weaver, Kevin Scott	0	0	3	0		0.75
677	ALUMBAUGH, Lonnie	0	0	5	1		1.5
678	PETTIT, Michael E.	2	4	1	1		2
679	SPENCER, Kelvin Demar	3	2	4	2		2.75
680	FLOYD, Lewis O.	2	0	1	2		1.25
681	ERVIN, Derrick	1	3	0	2		1.5
682	DAVIS, Rodney D.	1	3	1	1		1.5
683	TURRENTINE, Julius Lee	1	4	0	2		1.75
684	YOEUTH, Yoeun	1	3	1	1		1.5
685	PINA-ABOITE, Martin	1	3	0	1		1.25
686	WOOD, Terry L.	1	1	2	2		1.5

Figure 3: Screenshot of the Suspicion Analyzer with suspicion determinations scored on the far right.

The threshold number that we will have decided on will be based upon the total possible number of factors present, and the legal weight that they carry in court. For example, nervousness does not carry a heavy value because nearly everyone who is pulled over is nervous. Therefore, sweating and shaking hands may be scored at 1 or 2 respectively. However, factors like heavy odors of masking agents will be scored higher, at 4 or 5 because although they are not indicative, typically the heavy odor of masking agents is used by drug traffickers in an attempt to throw off the smell of a drug dog. This factor is more highly attributable to criminal activity than nervousness.

Looking forward in development the system will include prompts that will force the user of the system to further describe what they are seeing. For example, if nervousness is described, the user will be prompted to add why they believe a motorist is nervous and what degree of nervous they believe a motorist to be. This is significant for many reasons. A user may over or under describe the factors upon which they are relying in order to determine the suspicion, the computer will be programmed to a controlled variable of what suspicion is based on relevant case law. A separate algorithm will then correct the users answers if they are over or under representing what they see. For example, if an officer always says that a motorist is extremely nervous, the system will understand that that is the typical perception of the officer and will correct his error and place less of an emphasis on the officer's perception based on the systems understanding that this factor is always over exaggerated by that particular user.

We are well aware of the significant issue that bias, especially the misplaced racial bias plays in making determinations of suspicion. Therefore, our system will be designed to catch and eliminate bias at many stages. There are a few proposals considered on how to achieve this task. First, officers will be given their own identification number in the system, and their data on suspicion will be tracked separately as stated above. The computer will monitor the factors upon which officers rely. Race will also be tracked and will *not* be scored in anyway in the computer making the analysis of suspicion. Race may only be denoted to track whether or not a specific officer may have racial motivations in their decision making. For example, tracking race will allow us to see whether or not an officer on average finds suspicion for minorities, and

whether or not an officer may be entering data into the system falsely in order to achieve suspicion. The counter to this is that most officers wear a body camera and will be able to record what is happening, so in the rare scenario that officers seem to be racially profiling, camera footage may just prove that have been consistently at the wrong place at the wrong time dealing with similar motorists.

6. Identifying and Minimizing the Impact of Implicit Bias

While developing this project, we have noticed a currently unexplained phenomenon that we are currently investigating. While annotating opinions, it became quite clear that certain language gave the reader a feeling that the person being described was a minority. Although the race of the defendant was not given, the words themselves describing their actions and what the officers noticed about them seemed to convey that they were a racial minority. Upon further investigation by researching the defendant their race was discovered and many times we were correct in guessing their race simply based on the text that their actions had been described. Compare the two following sentences: "First, he noticed that appellant wore 'a lot of cologne.' The deputy described it as being 'a very overwhelming smell of cologne' and 'more than most people' would wear" and "The trooper also testified that he smelled a 'strong odor' of air freshener after he went up to the window and started speaking to appellant." As you can see here, each portion of text seems to be describing a factor that each officer is relying upon. However, upon further inspection we can tell that this is racially charged. Sentence one is describing a Hispanic man, he has been described as "more than most" people. However, the white person, in sentence two, had no comment about him compared to others, just the salient factor was described.

We are currently attempting to devise a mechanism in order to first, identify implicit racism and second a way to "sanitize" the language that is found. Much has been done in the way of "sanitizing" language that is implicitly biased. However, there has been considerable work done previously and there are different approaches to sanitize implicitly biased language. One such avenue to sanitize this language has been proposed by Giovanni Sileno, Alexander Boer and Tom Van Engers, who in *The Role of Normware in Trustworthy and Explainable AI*, discuss the potentially destructive nature of contemporary technology in society. The proposal there is the usage of oracles, specifically a second-order oracle which would be "specified by a neutrality constraint [that] would instead promote mutations of the source training data that satisfy it, e.g., pruning the data introducing the bias, or adding additional data to neutralize it." [6] However, it has been suggested that well-designed algorithms are capable of avoiding many cognitive biases. It has been posited that algorithms can be easily designed to avoid racial discrimination. [7]

For now, our approach has been to take a combination of these two theories. First, we are attempting to weed out the potential bias in our data by neutralizing or sanitizing our data from the start. As I explained, the data that we have noticed may be implicitly biased seems to compare people to people, presumably minorities to whites. We will sanitize our data by developing algorithms that specifically exclude any comparative racial language. A sanitizing algorithm could be placed after marker sets described above are created in order to filter out any comparative or racially biased language by specifically excluding words compiled in separate marker sets, which will be created similarly to the two corpora of data above.

7. Conclusion

Implementation of the first pilot programs, likely to be underway at the time of publication of this paper, will provide considerably more training data on the likelihood that drugs are actually present. At present, however, our prototype has demonstrated the potential of a more rational assessment of suspicion than is currently possibly by police or judges using case law. As odd calculations about the presence of drugs becomes possible, the prototype has the potential to legal definition of suspicion itself Finally, the prototype has the potential of addressing one of the most pressing concerns with drug interdiction stops – racial profiling.

REFERENCES

[1] Brinegar v. United States, 338 U.S. 160 (1949).
[2] Terry v. Ohio, 392 U.S. 1 (1968)
[3] D.H. Kaye, 2010. *The Double Helix and the Law of Evidence*. Harvard, Cambridge, MA.
[4] Noah Kupferbeg. 2008. Transparency: A new role for police consent decrees. Colum. J.L. & Soc. Probs. 42(1), 129-175.
[5] Damato v. State, 64 P.3d 700, 702 (Wyo. 2003)
[6] Giovannia Sileno, Alexander Boer, and Tom Van Engers. 2018. The Role of Normware in Trustworthy and Explainable AI. arXiv:1812.02471.
[7] Sunstein, Cass R., Algorithms, Correcting Biases (December 12, 2018). Forthcoming, Social Research. Available at SSRN: https://ssrn.com/abstract=3300171