ABSTRACT
A goal of software engineering research is advancing software quality and the success of the software engineering process. However, while recent studies have demonstrated a new kind of defect in software related to its ability to operate in fair and unbiased manner, software engineering has not yet wholeheartedly tackled these new kinds of defects, thus leaving software vulnerable. This paper outlines a vision for how software engineering research can help reduce fairness defects and represents a call to action by the software engineering research community to reify that vision. Modern software is riddled with examples of biased behavior, from automated translation injecting gender stereotypes, to vision systems failing to see faces of certain races, to the US criminal justice system relying on biased computational assessments of crime recidivism. While systems may learn bias from biased data, bias can also emerge from ambiguous or incomplete requirement specification, poor design, implementation bugs, and unintended component interactions. We argue that software fairness is analogous to software quality, and that numerous software engineering challenges in the areas of requirements, specification, design, testing, and verification need to be tackled to solve this problem.

ACM Reference format:

1 BIAS IN SOFTWARE
Bias is rampant in modern software. Try it out: type “He is a nurse. She is a doctor.” into http://translate.google.com and translate it into Turkish. Then translate the result (“O bir bebek hemşire. O bir doktor.”) into English and you get “She is a nurse. He is a doctor.” [18] (Figure 1). Want to upload a video to YouTube and have closed captions generated automatically? If your video features a man’s voice, the closed captions generated automatically? If your video features a man’s voice, the closed captions are likely to be more accurate than one with a woman’s voice [65]. Want your futuristic house or autonomous car to recognize you when you get in? Facial recognition systems often perform poorly on female and African American faces [40].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ESEC/FSE 2018–9 November, 2018Lake Buena Vista, FL, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnn.nnnnnn

Figure 1: Automatically translating text from a language without gendered pronouns into English can be influenced by societal biases [18].

for a job? Hopefully, when an employer searches for your name, the search engine doesn’t decide to show ads about your (nonexistent) arrest record only because of your race [64] (Figure 2). Evidence of bias in software goes on and on. In 2016, Amazon software decided not to offer same-day delivery to predominantly minority neighborhoods [43]. Staples offered online discounts to customers only in wealthier and more affluent neighborhoods [30, 52]. Language processing tools work better on English written by white people than people of other races [8]. The software US courts use to assess the risk of a criminal repeating a crime when deciding on bail and sentencing has been shown to exhibit racial bias [4].

With the increased use of machine learning in business, biases in underlying data used to train such systems can infiltrate the systems’ outcomes. The machine learning community is aware of this problem and has begun research on building machine classifiers that are fair in the face of biased data [16, 32, 73, 75], e.g., as part of its Fairness, Accountability, and Transparency in Machine Learning workshops (https://www.fatml.org/). The recently formed Conference on Fairness, Accountability, and Transparency (https://fatconference.org/) brought together, in 2018, researchers

Figure 2: Though since fixed, in 2013, Google searches for traditionally African American names (e.g., “Latanya Sweeney”) were more likely to result in ads for arrest records [64].
and practitioners interested in related topics. But while this conference was heavily attended by machine learning, policy, and legal experts, notably, software researchers and engineers were not well represented.

Despite the relatively low involvement, to date, of the software engineering community in this endeavor (with a few notable exceptions [2, 28, 68] discussed in Sections 2.3 and 2.4), the problems of producing fair software are, in fact, squarely in the software engineering domain. Much like other important software properties, such as quality or security, producing fair software requires (1) eliciting and specifying requirements that capture fairness properties, (2) architecting and designing systems while adhering to their fairness concerns, (3) validating and verifying fairness properties of the resulting software products, and (4) maintaining the fairness properties as the software evolves. As with software security and quality, it is insufficient to treat software fairness as an after-the-system-is-built concern; instead, fairness needs to be a first-class entity in the software engineering process, and doing so requires language, tool, and automation support at the levels of requirements, design, implementation, testing, and verification.

This paper represents a call to action for software engineering researchers to tackle the important challenges of engineering fair software. We outline the challenges that must be tackled to support developing fair systems and argue for the urgency of such research. Section 2 lays out the relevant challenges and presents research ideas our community should tackle. Section 3 places our call to action in the context of ongoing research. Finally, Section 4 summarizes our key arguments for the proposed line of research.

## 2. FAIRNESS RESEARCH CHALLENGES

How do we fight bias in software? Developing better machine learning methods is a start, but because there are many causes of bias, it cannot be the entire solution. Software fairness is analogous to software quality and security: good design and proper algorithms are important, but so are quality control via, e.g., testing and formal verification. For example, today, both companies and customers can use software testing to measure software quality, identify and report bugs, and prevent regressions. Similarly, tools are needed to support software fairness testing to measure software discrimination, enable identifying and reporting discrimination bugs, and protect against discrimination regressions — changes to code or data that introduce discrimination.

Experience has shown that high-level software concerns, such as security, quality, maintainability, self-adaptivity, etc. must be treated as first-class entities throughout the development lifecycle — not ignored until after the system is implemented — or else they are unlikely to be satisfied in the final product, e.g., [3, 19]. Fairness is similar, and when fairness is not made a first-class concern, systems result in unexpected, biased behavior, e.g., [18, 28, 40, 65]. This section outlines the research challenges that need to be addressed for software engineers to treat fairness as a first-class concern when building systems. This list is, of course, inherently partial, and more challenges are likely to arise as research progresses.

### 2.1 Requirements and Specifications

What does it mean for software to be fair? There have emerged numerous definitions of algorithmic fairness [55]. But while each of these definitions is appropriate in a given context, many are impossible to satisfy simultaneously [27]. For example, consider software that decides if an applicant should receive a loan, and consider fairness metrics with respect to race. The group fairness definition [16, 22, 32, 73, 75] states that the same fraction of the applicants of each race should get loans. If a higher fraction of applicants of one race receive loans than of another, that’s bias. Meanwhile the individual or causal fairness definition [22, 28] states that no two applicants identical in every way except race should result in one receiving the loan and the other not. So if there are two applicants who differ only in race, they should either both get a loan or neither of them should get a loan. Both of these properties seem desirable in a fair system. However, when applied to the real world, the two definitions cannot be (non-degeneratively, e.g., giving loans to no one) satisfied simultaneously. Suppose there is strong correlation between the applicants’ race and income (or any other attribute). Then it is not possible to both give the same fraction of individuals of all races a loan and to give all pairs of individuals identical in every way except race a loan. If only everyone above a certain income receives the loan, individual fairness is satisfied but group fairness is not. If the same fraction of applicants of each race receives a loan, then for the race correlated with lower incomes, some individuals who receive a loan must have lower incomes than some individuals of another race who do not receive a loan, and thus individual fairness cannot be satisfied.

The creation of a catalog of fairness definitions and guidelines for when each definition should be used is an open challenge in the engineering of fair software. Of the dozens of existing definitions [55], each depends on and is appropriate for certain contexts. To properly elicit and specify fairness requirements, the software engineers and often their clients must understand these contexts and the relationships between the contexts and the potentially desired fairness properties. Guidelines are needed to drive the questions clients should be asked, concerns and assumptions that need to be considered in selecting fairness definitions to use in specifying requirements, and the implications and restrictions imposed by these requirements.

The creation of requirement consistency and implication analyses is another open challenge. As mentioned earlier, combinations of fairness requirements that rely on multiple definitions of fairness may be, at times counterintuitively, mutually exclusive [27], and automated analyses or simulation can identify inconsistent or unsatisfiable requirements. Further, certain fairness requirements may result in unexpected or undesirable behavior. For example, recent work has shown that on one dataset, placing a constraint on a decision-tree machine-learning system [32] not to discriminate against gender significantly increased its discrimination against race [28]. Such an outcome is likely to be an unwelcome surprise to the software stakeholders who specified the requirement of no bias against gender. Analyses that help understand the possible implications of such constraints and how fairness requirements affect other requirements are critical for understanding the trade-offs and for properly specifying such requirements.
Further, analyses can help understand when satisfying a require-
ment does not achieve the overall desired goal. For example, the
group fairness property described above can be tricked to appear
as if the system technically satisfies the property, while failing to
satisfy it in spirit. Consider an instance of the loan recommendation
system that recommends loans to all tall applicants of race $\alpha$ but
to no short applicants of race $\alpha$, and to all short applicants of race $\beta$
but to no tall applicants of race $\beta$. Despite clear bias with respect
to race and height, a requirement of no group discrimination with
respect to race is satisfied because discrimination against short
applicants of race $\alpha$ is cancelled out by discrimination against tall
applicants of race $\beta$. This phenomenon can easily be missed when
specifying requirements, and automated analyses are needed to
mitigate such risks.

2.2 Architecture and Design

Inconsistencies among desired system properties are common. For
example, security and usability properties are often at odds, with se-
curity making systems harder to use. It is often difficult to tell prior
to developing a system architecture or prototype how requirements
will interact [62], and the case is likely to be no different for poten-
tially conflicting fairness requirements. An open research challenge
is to create tools that help model system architecture and identify
potential conflicts or trade-offs. Similar challenges in understanding
trade-offs, in terms of functional and non-functional properties of
systems, have been tackled with discrete-event simulation [61, 63]
and fuzzy logic [23], and such approaches can potentially help with
making fairness-related design decisions. Multi-objective optimization
can help identify fair trade-offs when considering treating
multiple software stakeholders fairly when eliciting requirements
from them [24]. Similar approaches could perhaps be used to bal-
ance mutually opposing fairness requirements.

Years of system-building experience have produced architectural
styles and design patterns for various desirable software proper-
ties [66], e.g., the peer-to-peer architectural style to avoid single
points of failure. An open research challenge is to develop styles
and patterns for fairness properties, perhaps augmenting the abovemen-
tioned catalog of fairness definitions with which styles and
patterns help address each definition. Further, style and pattern
interactions need to be studied to understand which ones can, and
which ones cannot be composed.

For systems that rely on machine learning, the design of fairness-
aware machine learning algorithms that can produce fair models
even when faced with biased training data is critical. Such algo-

drithms are in relatively early stages of development, e.g., [16, 32,
73, 75], but more mature frameworks with theoretical guarantees
are emerging [67].

2.3 Testing and Debugging

The primary method for ensuring software quality is testing [3].
There is strong reason to believe the same should be true for soft-
ware fairness. Even advanced tools to ensure fairness of require-
ments and design cannot prevent all bugs, unintended interactions,
assumptions that fail to hold in the real world, and other causes of
bias. Testing can, as with quality, assess the fairness of the built
system, provide a mechanism for debugging bias, and guard against
bias regressions.

To date, relatively little software engineering research has fo-
cused on these issues, with the notable exceptions of Themis [287 ]
and FairTest [68]. Themis is a mechanism for automated test
generation for systems with categorical inputs, such as, for exam-
ple, the loan recommendation system described above. Themis
automatically generates fairness test suites for group and causal
(individual) fairness; executing these test suites produces a fairness
score and a measure of confidence in that score. FairTest allows
developers to write manual tests and performs various analyses to
measure fairness-related statistics.

Fairness bugs are common in software systems with complex
inputs and outputs, to which Themis and FairTest cannot directly
apply. For example, tools with natural English inputs parse Eng-
lish written by white people more accurately than that written
by people of other races [8], and facial detection and recognition
tools’ accuracy also depends on demographic information, such as
race and gender [40]. A major challenge for automated fairness
test generation is to generate tests for systems with such inputs.
Testing such systems for causal fairness requires the generation of
carefully controlled natural language and photos that are identical
except for a fixed set of sensitive attributes. Further, automated test
generation does not yet address covariants in attributes; accounting
for such covariants can generate more accurate test suites.

The large feature spaces involved in applications with potential
bias create a significant scaling challenge for testing. Exhaustive
testing is, of course, infeasible for real-world systems, but some
definitions of discrimination, including causal discrimination, are
susceptible to provably-sound pruning and adaptive sampling to
reduce the test space [28]. Further efficiency improvements and
guided test generation are needed to make fairness testing applicable
to large, real-world systems. For example, new methods may
exploit discovered tests that identify evidence of bias to guide sub-
sequent test generation and may be able to find more evidence faster
than random search. Further, fairness testing requires executing
the software under test many times, often with similar inputs. This
offers an opportunity for improving the efficiency of execution
using information from prior executions. This incremental execu-
tion optimization could, potentially, greatly reduce test execution
time and increase the applicability of fairness testing to larger sys-
tems. Similarly, test prioritization and selection may improve the
efficiency of fairness testing systems.

As with software quality, while testing can identify fairness bugs,
developers also need debugging tools to understand and remove the
root causes of the bias. Root causes can lie in the involved require-
ments, algorithms, implementation, or data [7], and an extensive
suite of debugging tools is necessary to support the developers,
including tools for debugging how data inaccuracies may affect
software outcomes. As an example of data imperfections being a
root cause of bias, such imperfections may affect certain sensitive
groups more than others: Due to database errors, which are more
prevalent in DHS records than SSA records, the E-Verify program
(the voluntary, government-run system that employers can use
to check whether new employees are work-eligible) had rejection
rates 30 times higher for naturalized citizens and 50 times higher for
legal nonimmigrants, than for natural-born citizens [59]. Ideas from research on inferring causal models [44] and relationships [48, 50] in data can be adapted to build models of causal relationships in software behavior, aiding system understanding and debugging tasks. Such fairness analysis has the potential to produce powerful tools because it combines the ability to obtain new execution data on demand by running the software under analysis, and to conduct causal experiments to test causal hypotheses [28].

2.4 Verification

As with all software correctness, verification is a highly desirable goal for software fairness. Recent work proposing verification of fairness properties highlights relevant challenges [2]. First, fairness properties are often aggregates of many executions, as opposed to correctness properties that can be invalidated by a single counterexample execution. This can complicate model-checking and even make checking assertions more challenging. Second, fairness properties are often not only reflective of the software but also of the input population relevant to the software execution, making it more difficult, or less relevant, to verify the software for all populations. Third, fairness properties are often probabilistic, restricting which existing verification techniques can be applied directly to the problem.

Another research challenge with respect to formal verification of fairness properties is identifying ways to encode fairness definitions as verifiable program properties. Some such definitions can be encoded as probabilistic program properties [2], though others may require different mechanisms. Developing runtime environment support for asserting these properties during execution can also help ensure proper behavior.

Finally, support for debugging fairness bugs when formal verification finds behavioral counterexamples or evidence of bias is needed to help developers not only detect bias but also remove it and improve systems.

3 RELATED WORK

Discrimination shows up in many software applications, e.g., advertisements [64], hotel bookings [45], and image search [36]. At the same time, software is entering domains in which discrimination could result in serious negative consequences, including criminal justice [4], finance [56], and hiring [60]. Software discrimination may occur unintentionally, e.g., as a result of implementation bugs, as an unintended property of self-organizing systems [9, 11, 13, 14], as an emergent property of component interaction [10, 12, 15, 41], or as an automatically learned property from biased data [16, 17, 31–34, 73–75].

Existing work on test generation focuses on two specific measures of discrimination, group and causal. Group discrimination is a generalization of the Calderas-Verwer (CV) score [17], used frequently in prior work on algorithmic fairness, particularly in the context of fair machine learning [16, 32, 73, 75]. Many other definitions exist [55]. One defines discrimination by observing that a “better” input is never deprived of the “better” output [22]. That definition requires a domain expert to create a distance function for comparing inputs. Causal discrimination [28] measures causality [58]. Fairness in machine learning research is also concerned with causal measures of discrimination, e.g., counterfactual fairness [42]. FairML [1] uses orthogonal projection to co-perturb characteristics, which can mask some discrimination, but find discrimination that is more likely to be observed in real-world scenarios.

FairTest [68] uses manually written tests to measure four kinds of discrimination scores: the CV score and a related ratio, mutual information, Pearson correlation, and a regression between the output and sensitive inputs.

Reducing discrimination in machine learning classifiers [16, 32, 67, 73, 75] is important work that is complementary to the effort the software engineering community needs to put in to create tools and other support for developing fair software.

Some fairness properties can be encoded as probabilistic program properties over a distribution of executions and asserted at runtime or even formally verified with probabilistic guarantees [2]. This work can form the basis of future research on formally verifying fairness properties.

Combinatorial testing may be relevant to fairness testing as it aims to minimize the number of tests needed to explore certain combinations of input characteristics. For example, all-pairs testing generates tests that evaluate every possible value combination for every pair of input characteristics, which can be particularly helpful when testing software product lines [5, 35, 37, 38]. The number of tests needed to evaluate every possible value pair can be significantly smaller than the exhaustive testing alternative since each test can simultaneously contribute to multiple value pairs [20, 35, 69]. Such combinatorial testing optimizations may aid fairness testing. Advances in combinatorial testing, e.g., using static or dynamic analyses for vacuity testing [6, 29] or to identify configuration options that cannot affect a test’s output [59], can directly improve efficiency of discrimination testing by identifying that changing a particular input characteristic cannot affect a particular test’s output, and thus no causal discrimination is possible with respect to that particular input.

It is possible to detect discrimination in software without explicit access to it. For example, AdFisher [21] collects information on how changes in Google ad settings and prior visited webpages affect the ads Google serves. AdFisher computes a variant of group discrimination, but could be adapted to measure other fairness requirements as well.

Causal relationships in data management systems [26, 46, 47] can help explain query results [51] and debug errors [70–72] by tracking and using data provenance [49]. For software systems that use data management, such provenance-based reasoning may aid testing for causal relationships between input attributes and outputs. Our prior work on testing software that relies on data management systems has focused on data errors [53, 54], whereas this work focuses on testing fairness.

Automated test generation tools, e.g., Randoop [57] and EvoSuite [25] target improving software quality, but their underlying mechanisms may be helpful for generating fairness test suites as well. Fairness testing, however, involves more than generating test suites. It also processes the results of these tests to compute bias. Some generation can use results of the analyses in an iterative loop to determine when to generate more tests and to guide which tests to generate. The goals of traditional test generation tools also differ...
Software Fairness

from the goal of fairness testing, aiming instead to generate tests for testing goals, e.g., maximizing coverage, which leads to diverse test suites [25, 57]. Fairness testing often needs to generate pairs of similar, not diverse, inputs that traditional tools are unlikely to generate.

4 SUMMARY

Software fairness is an increasingly important problem for today’s software that should be tackled from multiple directions. In particular, ensuring software fairness is a software engineering problem, with deep apparent parallels between software fairness and software quality. Requirements elicitation, design, testing, and verification – all critical parts of the software engineering lifecycle and of ensuring software quality – are all relevant and necessary for ensuring software fairness, but significant research challenges exist in creating developer support in each of these areas.

There are early encouraging signs that the software engineering community is getting involved in software fairness, such as the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare 2018). Combined with the machine learning community’s efforts, this challenging problem provides numerous opportunities for interdisciplinary, collaborative solutions. The potential for significant societal impact and many challenging and interesting research problems, such as fairness requirements specification, automated test generation, and efficient test selection and execution make these exciting research problems worthy of the attention of software engineering researchers, as well as researchers in data science, machine learning, theoretical computer science, and other disciplines.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation under grants no. CCF-1453474, IIS-1453543, and CNS-1744471.

REFERENCES


Dana Mattioli. On Orbitz, Mac users steered to pricier hotels.


Aarti Shahani. Now algorithms are deciding whom to hire, based on voice. NPR All Things Considered, 2015.


