A Study on the Prevalence of Human Values in Software Engineering Publications, 2015 – 2018

Anonymous Author(s)

ABSTRACT

Failure to account for human values in software (e.g., equality and fairness) can result in user dissatisfaction and negative socio-economic impact. Engineering these values in software, however, requires technical and methodological support throughout the development life cycle. This paper investigates to what extent top Software Engineering (SE) conferences and journals have included research on human values in SE. We investigate the prevalence of human values in recent (2015 – 2018) publications in these top venues. We classify these publications, based on their relevance to different values, against a widely used value structure adopted from the social sciences. Our results show that: (a) only a small proportion of the publications directly consider values, classified as *directly relevant publications*; (b) for the majority of the values, very few or no directly relevant publications were found; and (c) the prevalence of directly relevant publications was higher in SE conferences compared to SE journals. This paper shares these and other insights that may motivate future research on human values in software engineering.

KEYWORDS

Human Values, Software Engineering, Paper Classification

1 INTRODUCTION

Ignoring human values while engineering software may result in violating those values [13, 25] and subsequent dissatisfaction of users. This may lead to negative socio-economic impacts such as financial loss and reputational damage. A recent example, which made news headlines, is the price gouging on airline tickets during Hurricane Irma [29]. After a mandatory evacuation order, the cost of airline tickets rose six fold, due to supply and demand pricing systems, thus disadvantaging evacuees. Arguably, this occurred because of insufficient consideration of valuing compassion for those suffering in a natural disaster. A second example is software used by Amazon to determine free shipping by zip code, which turned out to discriminate against minority neighbourhoods [19]. The COMPAS system, used by US parole boards to predict re-offenders, has been shown to suffer from racial bias [2]. Indeed, the negative impacts of ignoring values can go as far as risking human life: the tragic suicide of the British teenager Molly Russell [4] has been partially attributed to Instagram's personalisation algorithms, which flooded Molly's feed with self harm images. Following public outrage, Instagram has now banned such images.

As awareness about human aspects of software grows, the public is increasingly demanding software that accounts for their values. See, for example, those accusing Facebook of taking advantage of users' data to influence the US elections [39]. Public demand has also motivated software vendors to take preemptive measures to avoid violating human values. Google, for instance, has pledged not to use its AI tools for surveillance conflicting with human rights [9].

Though such initiatives are promising, we question whether software engineering research and practice currently pays sufficient attention to human values. Whilst some values (such as privacy, security, and accessibility) are well embedded in SE methods, others (such as integrity, compassion, and social justice) have received less attention. This may be due to the lack of adequate methodological and technical support for engineering some kinds of values in software [25].

In this paper, we investigate to what extent research in top-tier SE conferences and journals has considered the full range of human values. In particular, we classified publications in the International Conference on Software Engineering (ICSE), the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), the IEEE Transactions on Software Engineering (TSE), and the ACM Transactions on Software Engineering and Methodology (TOSEM) from 2015 to 2018, based on their relevance to different values. Whilst we acknowledge that other (sub-)disciplines, such as HCI or Requirements Engineering, may contain work on SE and human values, classifying such venues is out of the scope of this paper and would warrant further study. This paper addresses whether those venues generally considered to be top general SE venues address human values. If the top venues do not address values, then the SE field may want to find ways to lift the prominence of research on values in SE.

The paper addresses the following three research questions:

- **(RQ1)** To what extent do publications in top SE venues address human values in software?
- **(RQ2)** Which values are commonly considered in publications in top SE venues?
- **(RQ3)** How are those publications addressing human values distributed across venues?

Our approach was to read the abstracts of papers in ICSE, FSE, TSE and TOSEM from 2015-2018 (1350 papers in total) and manually classify each paper (using multiple raters) as either "directly relevant" or "not directly relevant" to one or more human values. Classification based on abstracts is a well-accepted method in the SE literature [37, 43]. For a definition of human values, we used a widely adopted value framework (Figure 1), Schwartz's theory of human values [30, 35], which is well-accepted in the social sciences and defines 58 human values. A paper was classified as *directly relevant* to a particular value if its main research focus is to define, refine, measure, or validate a particular value, or propose a solution (e.g., a tool, technique or methodology) to address one or more human values in software. We use "direct relevance" rather than simply "relevance" because papers often make broad statements in their introductory text about the benefit of the work to society but such claims may not be the main research focus nor have been validated. "Direct relevance" was assigned only to those papers with a main focus on human values. Multiple raters were used to come to consensus on the classification. There is inevitably, however, some degree of subjectivity in the classification. We therefore followed recommended practice in social science to mitigate threats to validity.

The results of our study showed that: (a) only 16% of publications were directly relevant to human values; (b) for 60% of human values, there were no directly relevant publications; (c) for 79% of human values, the number of directly relevant publications was ≤ 2; only 21% of values had on average 2 directly relevant publications, and (d) 88% of directly relevant publications were found in SE conferences rather than journals.

2 BACKGROUND

Cheng and Fleischmann summarize seven different definitions of human values as “guiding principles of what people consider important in life” [8]. Human values with an ethical and moral import such as equality, privacy and fairness have been studied in technology design and HCI for more than two decades [15–17]. Meanwhile, the rapid popularization of artificial intelligence (AI) and its potential negative impact on society have raised the awareness of human values in AI research [7, 11, 27]. Consequently, human values are getting renewed research focus.

There has been some recent (but isolated) research attempts in SE related to human values such as values-based requirements engineering [44], Values-First SE [13] and Values-Sensitive Software Development [1]. However, there has been no previous work that measures to what extent human values have been considered in SE research. Motivated by this research gap, we follow a classification approach, similar to that used in previous SE research to map topic trends [23, 37, 42], but with a different purpose, to measure values relevance. There are no current classification schemes for human values in SE. Therefore, we take inspiration from the social sciences.

Social scientists have been searching for the most useful way to conceptualize basic human values since the 1950s [34]. In 1973, Rockeach captured 36 human values and organized them into 2 categories [28]. In 1992, Schwartz introduced his theory of basic human values (henceforth referred to as Schwartz’s Values Structure (SVS)) which recognized 58 human values grouped into 10 value categories [30, 32]. While these two value structures remain the most well recognized ways of representing values, there are at least ten other value classifications [8]. In this paper, we use Schwartz’s theory, which is the most cited and most widely applied classification in the social sciences [44]. It has also been applied in numerous computer science [5, 26] and SE studies [14]. For example, Schwartz was used to incorporate values in the SE decision making process [13], to measure the values of software developers [49] and to predict movie genres for certain personality types [26].

In SVS, Schwartz introduced 10 motivationally-distinct value categories recognized across more than 20 cultures [30]. Each value category has underlying distinct motivational goals (see Table 1), which relate to three fundamental needs of human existence [30]. Schwartz subdivided each value category into a set of closely related values [30, 31]. These 10 value categories and 58 values are arranged in a circular motivational structure as shown in Figure 1. Value categories located close to each other are complementary whereas those further apart tend to be in tension with each other. Section 3 discusses how we applied SVS in our classification study.

Table 1: Value categories and descriptions [35]

Value Category	Description (motivational goals)
Self-direction	Independent thought and action–choosing, creating, exploring
Stimulation	Excitement, novelty, and challenge in life
Hedonism	Pleasure or sensuous gratification for oneself
Achievement	Personal success through demonstrating competence according to social standards
Power	Social status and prestige, control or dominance over people and resources
Security	Safety, harmony, and stability of society, of relationships, and of self
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms
Tradition	Respect, commitment, and acceptance of the customs and ideas that one’s culture or religion provides
Benevolence	Preserving and enhancing the welfare of those with whom one is in frequent personal contact
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature
Holistic View	Human values considered holistically without focusing on predetermined values

3 METHODOLOGY

We manually classified publications from SE conferences and journals, generally considered to be the top general SE venues. The aim was to assess the prevalence of human values in publications in these leading venues.

Leading venues were identified as ICSE, ESEC/FSE, TSE, and the TOSEM. These venues are historically accepted in the SE community as the top two general SE conferences and journals; this is also backed up by metrics (e.g., guide2research which rates ICSE and ESEC/FSE as the top SE conferences based on h-index, and Robert Feldt’s journal ranking [12] which has TSE as 1st and TOSEM as 3rd – Empirical Software Engineering is 2nd but is a more specialist journal and so was not included in our classification).

To classify papers against values, we followed a methodology similar to that of prior classification work in SE [23, 37, 42]. As with prior studies, ours was based on manual classification of paper abstracts by multiple raters. Classification based on abstracts, rather than reading the full paper, is sub-optimal but strikes a balance between accuracy and time needed for the study. All papers had multiple raters and inter-rater agreement was measured using Fleiss’ Kappa [22]. In total, we employed seven raters (5M/2F) with varying levels of experience in SE research, ranging from PhD students to professors. Note that this is a relatively high number of raters compared to similar studies [6, 46].

When conducting such a study, there are a number of key experimental design decisions that need to be taken, including: (i) how

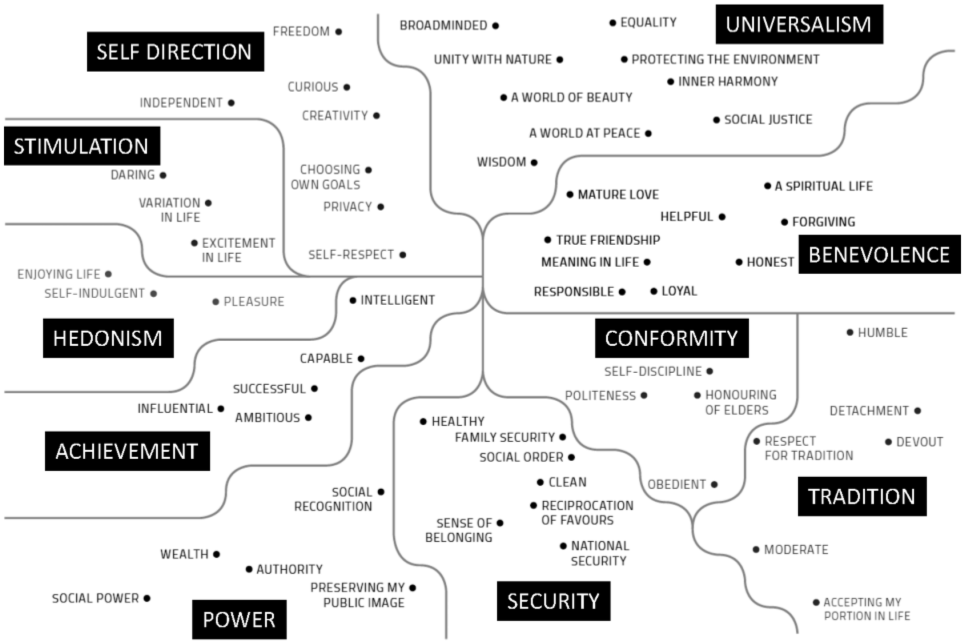

Figure 1: Schwartz Values Structure [33, 36] (adopted from [20]). Words in black boxes are values categories, each subdivided into values.

to define relevance to human values, given the imperfect and high-level nature of values definitions in the literature; (ii) how many raters to assign to each paper, and (iii) how to resolve disagreements between raters. To make choices about these design decisions, we first carried out a pilot study before carrying out the main study. Both the pilot and main study assumed SVS as the classification scheme. All raters had reasonable knowledge about SVS and had conducted research on socio-technical aspects in SE.

3.1 Pilot Study

The pilot study had three steps: (i) Paper selection and allocation of papers to raters, (ii) Paper classification, and (iii) Calibration of classification decisions made by different raters. The aim of the pilot study was not to measure relevance of papers to values; rather, we had the following objectives:

- To test the appropriateness of SVS as the classification scheme for SE publications
- To develop a common understanding regarding the meaning of human values in SE contexts
- To collect insights from raters to feed into the experimental design of the main study

(i) Paper selection and allocation of papers to raters. We randomly selected 49 papers from ICSE 2018 as our pilot study dataset. These were equally allocated among the seven raters, with three raters per paper. Common practice is to assign two raters per paper [6, 46]; three were assigned in the pilot to get a better understanding of how to map papers to values. ICSE was chosen as it has the broadest coverage of SE research [6]. We chose the most recent ICSE proceedings – 2018 at the time of the study.

(ii) Paper classification. Raters classified papers, independently, based on the title, abstract and keywords which is an approach used in similar classification studies in SE [6, 18, 37]. Raters were instructed to decide if a paper was "relevant" or "not relevant" to human values: relevance was deliberately left ill-defined as one of the objectives of the pilot was to influence the definition of this term in the main study. For relevant papers, raters were asked to classify the papers into one value category (e.g. *Power*), and then into one value within the category such as *Wealth* or *Authority* (see Figure 1). Raters were not mandated to follow the hierarchical structure of SVS: that is, they could classify a paper into value X and value category Y even if X did not belong to category Y. This was to give us a way to assess the appropriateness of the hierarchy in SVS.

(iii) Calibration. After classification, all seven raters met to discuss the classification decisions. The main objective was to calibrate decisions and use this to refine the definition of values relevance. The intention was *not* to decide which rater picked the correct classification.

Following the pilot study, we made a number of observations which were fed into experimental design of the main study.

- *Observation 1:* Raters found that almost every paper could be classified into a small number of values such as *Helpfulness*, *Wisdom* or *Influence* because, in general, all research tries to advance knowledge. Thus, an indirect argument could almost always be made why a paper is relevant to helpfulness (e.g., a paper on testing is helpful to testers), wisdom (any paper advances knowledge, thus leading to greater wisdom), or influence (e.g., a paper on an improved software process influences how software is developed). This observation illustrated the difficulty of working with

vaguely defined concepts such as values, but also the importance of a better definition of relevance.
Decision 1: It is beyond the scope of this paper to fully and formally define all the values; hence, it was decided in the main study to use inter-rater agreement as evidence that a value was sufficiently understood in the context of a particular paper to provide confidence in the results. The definition of relevance was, however, refined for the main study. Raters were instructed not to make indirect arguments why a paper might be relevant to a value. Instead, in the main study, classification was based on "direct relevance" – a paper is defined as directly relevant to a value if its research focus is to define, refine, measure, or validate a particular value or propose a solution (e.g. a tool, technique or methodology) aimed at addressing a human value. This revised definition places emphasis on those papers with a main research contribution of a particular value, not merely a broad statement about relevance to a value.

- *Observation 2:* Raters observed that some papers addressed values as a general concept rather than considering any specific value. An example would be a paper that presents a methodology for refining values into a software architecture. These papers should not be classified into any particular value category or value.
Decision 2: To facilitate classification of such papers, we introduced a new value category in the main study, named *Holistic View*. A paper classified under *Holistic View* relates to values generally without focusing on any specific value (Table 1).
- *Observation 3:* Raters found that some papers should be classified under more than one value.
Decision 3: To accommodate such papers in the main study, raters were allowed to select up to three values. This decision is different from similar studies in SE where raters were obliged to pick just one category [6].
- *Observation 4:* The pilot study gave us an opportunity to measure how long it took raters to rate papers. We found that, on average, each rater spent four minutes per abstract. Given the number of papers in the main study (1350 – see Table 2), assigning three raters per paper would be infeasible.
Decision 4: Out of necessity, we reduced the number of raters in the main study to two. This is consistent with the number of raters in similar studies [6, 18, 46].

3.2 Main Study

Similar to the pilot study, the main study also had three phases: (i) Paper selection and allocation of papers to raters, (ii) Paper classification and (iii) Disagreement resolution. The final stage was different to the pilot study because rather than calibrating ratings to inform experimental design, raters in disagreement met to try and reach a consensus.

(i) Paper selection and allocation of papers to raters. For the main study, we selected papers from ICSE, FSE, TSE and TOSEM over the last four years. These are the same venues used in similar paper classification studies [6, 18]. We selected all papers in TSE and TOSEM. For FSE, we used all papers from the main track, and for ICSE, we used all papers from the main track, from the Software Engineering in Practice (SEIP) track, and from the Software Engineering in Society (SEIS) track. SEIP was included to acknowledge the prominence of this industry-focused track at recent ICSEs. SEIS was included as it has a specific focus on social and societal aspects of software engineering. In total, there were 1350 papers published in the chosen venues over the years 2015–2018 (see Table 2). This is a high sample size compared to similar studies (e.g., 976 in Bertolino et al. [6] and 369 in Glass [18]). The papers were randomly allocated among the seven raters, two raters per paper. Each rater received around 400 papers to classify. We manually extracted links for each of the 1350 papers from digital databases, provided a spreadsheet with these links as well as values and value categories for raters to select from.

Table 2: Classified publications by venue/track and year

Venue & Track	2015	2016	2017	2018	Total
ICSE–Main Track	83	101	68	153	405
ICSE–SEIP	25	28	30	35	118
ICSE–SEIS	9	7	9	11	36
ESEC/FSE–Main Track	123	143	124	122	512
TSE	62	61	61	31	215
TOSEM	22	16	12	14	64
Total	324	356	304	366	1350

(ii) Paper classification. Similar to the pilot study, raters were asked to classify papers on the basis of the title, abstract and keywords. However, the main study used a different definition of relevance, as suggested by the pilot study. Raters were asked to classify papers as directly relevant or not directly relevant, where the definition of direct relevance is as given in Section 3.1. Papers found directly relevant to values were further classified into a category and then to a specific value(s). Throughout the process, raters complied with the decisions made during the *calibration* step in the pilot study.

(iii) Disagreement resolution. Given the subjective nature of the classification, raters sometimes disagreed. This could arise at three levels: (a) relevance level, where raters disagreed on whether a paper was directly relevant or not; (b) value category level, where raters disagreed on the choice of value category; and (c) value level, where raters disagreed on the choice of value.

To attempt to resolve these disagreements, raters met to discuss their views about why the paper in question was classified in a certain way. If the raters could not come to an agreement, a third rater was introduced as an arbiter. The arbiter facilitated a second round of discussion, sharing his or her own views, to facilitate a consensus. However, if the disagreement persisted, the arbiter did not force a decision.

Aligned with previous studies [6], we calculated inter-rater agreement using Fleiss' Kappa, once attempts at resolving disagreements had taken place. The results of the Kappa measure are interpreted according to the agreement strengths introduced by Landis and Koch [22]. We achieved *almost perfect* agreements on relevance level and category level with Kappa values equal to 0.92 and 0.87, respectively. The agreement of value level was found as *substantial* with Kappa value equal to 0.79. The results from the main study are further discussed in Section 4.

4 RESULTS

This section presents the results of the main study described in Section 3.2. As a reminder, we investigate the following research questions:

(RQ1) To what extent do publications in top SE venues address human values in software?

(RQ2) Which values are commonly considered in publications in top SE venues?

(RQ3) How are those publications addressing human values distributed across venues?

4.1 RQ1: human values prevalance in publications in top SE venues

Figure 2 demonstrates the prevalence of human values in classified publications. We observed (Figure 2) that the majority of the publications (82%) were classified as not directly relevant to Schwartz values, which constitutes 1105 out of 1350 papers.

Table 3 gives an example of a paper classified as not directly relevant (row 1) – the paper does not directly focus on addressing any particular Schwartz value. 16% of the publications (216 papers) were found to be directly relevant to values. The remaining 2% of publications (29 papers) were classified as undecided, because the raters could not agree on a classification.

To investigate if there were any trends in the prevalence of values in these SE venues over time, we compared the percentages of the directly relevant publications from 2015 to 2018 (Figure 3): no significant trends were observed.

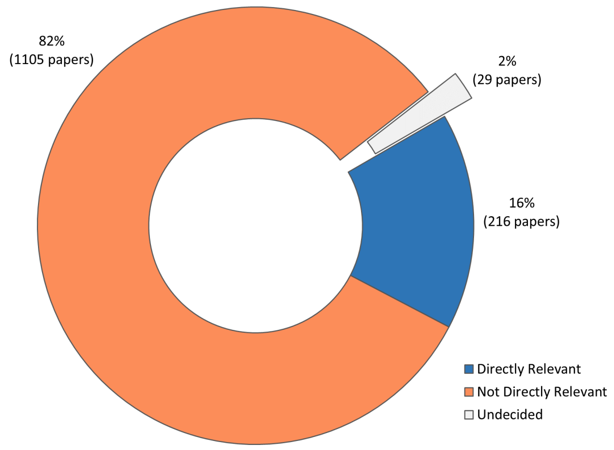

Figure 2: Relevance of SE publications to human values

4.2 RQ2: Which human values are most commonly considered?

Our results show that out of the 58 Schwartz values in Figure 1, the ones that were found in the publication sample of this study, had on average 2 directly relevant publications.

As shown in Figure 5, however, the frequency of the directly relevant publications varied significantly for different values. Figure 4 shows the level of attention given to the 58 human values in SVS. It can be seen that for the majority of the values (79%), the number of directly relevant publications was ≤ 2 while for 60% (35 out of 58) of the values, no directly relevant publications were found (Figure 4).

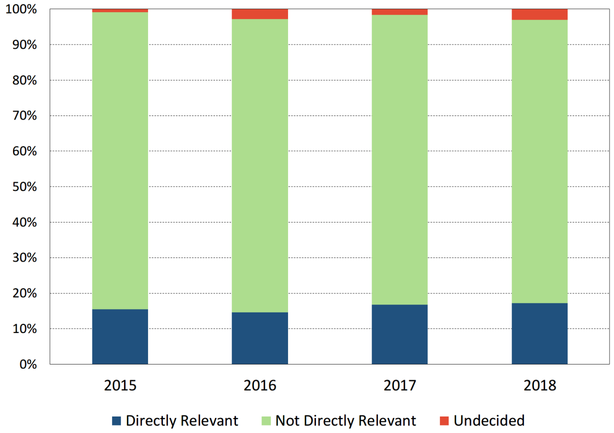

Figure 3: Publications per year

Also, for some values, e.g., *Enjoying Life* and *Honoring of Parents and Elders*, only one directly relevant publication was found (Figure 5). It can also be seen in Figure 4 that only for 21% (12 out of 58) of the values, e.g. Helpful and Privacy, the number of the relevant publications was above average (> 2). While being cautious with generalizing, these findings are highly suggestive of negligible or limited attention paid by the top SE research venues to the majority of human values.

In the attempt to understand which values are most commonly considered, we found (Figure 5) that the number of publications relevant to *Helpful*, *Privacy*, and *Protecting the Environment*, were the highest. Examples of such publications are given in Table 3. With 38 relevant papers, the value *Helpful* was the most frequently considered value. Publications that contributed software tools, techniques or methodologies developed to enhance the welfare of others were classified by the raters as directly relevant to the value *Helpful*.

The second highest number of directly relevant publications was observed for *Privacy* (Figure 5). This group contained papers that directly considered user privacy. Also, *Protecting the Environment*, the third most commonly found value, appeared in publications that directly considered sustainability and energy efficiency in software.

It can be observed from Figure 6 that 80 papers (41% of the relevant publications) were classified as directly relevant to *Security*, which made *Security* the most prevalent value category. This is not surprising as security is a well-recognized quality aspect of software, for which there is a great demand from stakeholders. The second and third most highly prevalent value categories were *Benevolence* and *Universalism*, which constituted 20% and 16% of the values publications, respectively. On the other hand, no publications were found to be relevant to the categories *Tradition*, *Stimulation*, and *Hedonism*. Moreover, 8% of the relevant papers were classified under the category *Holistic View*, which does not exist in SVS – this category was introduced based on the raters' feedback from the pilot study (Section 3.1) to account for publications that considered values in general.

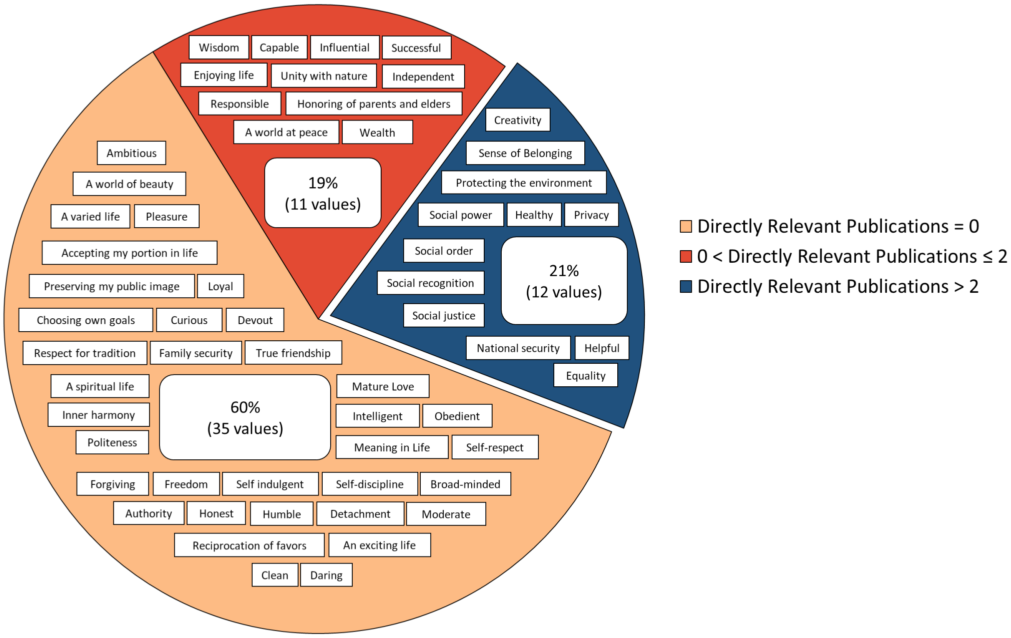

Figure 4: The level of attention given to 58 values in the Schwartz Value Structure

Table 3: Examples of paper classification at different levels (*direct relevance*, *value category*, and *value*)

Classification	Extracts from Abstract
Not Directly Relevant	…system calls provide us with a window into the development process and design decisions that are made for the Linux kernel …presents the result of an empirical study of the changes (8,770) that were made to the system calls during the last decade (i.e., from April 2005 to December 2014) …As of December 2014, 396 system calls existed in the Linux kernel. They can be categorized into 10 groups (process management, signal processing, and so on) …[3]
Privacy	Network traffic data contains a wealth of information for use in security analysis and application development. Unfortunately, it also usually contains confidential or otherwise sensitive information, …We present Privacy-Enhanced Filtering (PEF), a model-driven prototype framework that relies on declarative descriptions of protocols and a set of filter rules …[10]
Helpful	…However, newcomers face many barriers when making their first contribution to an OSS project, leading in many cases to dropouts. Therefore, a major challenge for OSS projects is to provide ways to support newcomers during their first contribution. In this paper, we propose and evaluate FLOSScoach, a portal created to support newcomers to OSS projects. …[40]
Protecting the Environment	…The battery power limitation of mobile devices has pushed developers and researchers to search for methods to improve the energy efficiency of mobile apps. We propose a multiobjective refactoring approach to automatically improve the architecture of mobile apps, while controlling for energy efficiency …[24]
Holistic View	…The aim of this paper is to give more visibility to the interrelationship between values and SE choices. To this end, we first introduce the concept of Values-First SE and reflect on its implications for software development. Our contribution to SE is embedding the principles of values research in the SE decision making process and extracting lessons learned from practice …[13]

4.3 RQ3: Differences between venues

To answer **RQ3**, this section reports our findings on the distribution of values-relevant publications across the four venues. Figure 7 demonstrates, for each venue/track, the proportion of the directly relevant publications in 2015-2018.

The proportion of directly relevant publications in each venue/track. We observed (Figure 7) that the proportion of directly relevant publications in the two SE journals, namely TOSEM (about 5%) and TSE (about 11%), is lower than the proportion in the main tracks of ICSE (about 18%) and FSE (about 13%), and significantly lower than the proportion in the SEIP (21%) and SEIS (about 81%) tracks of ICSE. In particular, the proportion of values-relevant papers was significantly higher in SEIS. This is not surprising given the focus of the track.

The distribution of directly relevant publications by venue/track. Figure 8 shows the distribution of relevant publications across the venues/tracks. From all 216 publications that directly considered values, 58% were published in different tracks of ICSE: main track (33%), SEIS (14%), and SEIP (11%). The highest prevalence of directly relevant publications was in the main tracks of ICSE (33%) and FSE (30%). As such, it was concluded that about 88% of the publications

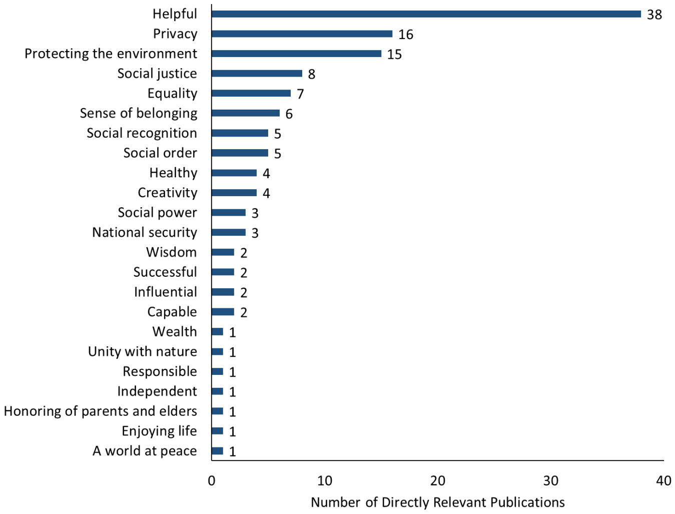

Figure 5: The number of directly relevant publications per value

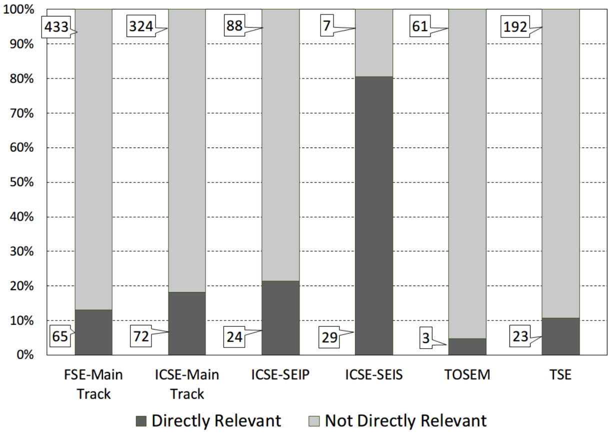

Figure 7: Differences in directly relevant publications across venues/tracks. Labels on the bars denote the number of papers in each category.

that directly considered values were published in the SE conferences: ICSE (58%) and FSE (30%). On the other hand, SE journals, TSE (11%) and TOSEM (1%), constituted only 12% of the directly relevant publications (Figure 8).

The distribution of directly relevant publications by values and venues. Figure 9 shows how the publications directly relevant to different values are distributed across different venues/tracks. We observed that only 23 out of 58 values in SVS were present. For some values, publications were found across most venues/tracks. For example, publications directly relevant to Helpful were found in 5 out of 6 venues/tracks. But for the majority of the values in Figure 9 (15 out of 23), the number of the venues/tracks that published papers for those values did not exceed 2. For instance, publications directly relevant to Social justice and National security were found only in the main tracks of FSE and ICSE. Also, publications relevant to Enjoying life, Honoring of the parents and elders, and A world at peace appeared only in the main track of ICSE. Publications for certain values, e.g., Equality, Social justice, and Healthy, were only present in conference papers but not in journals. We further observed that for the majority of values (19 of 23 values in Figure 9), relevant publications were found in the main track of ICSE while publications in TOSEM only considered Privacy.

The distribution of directly relevant publications by value categories and venues. Publications relevant to 7 out of 10 value categories in SVS were found across different venues/tracks (Figure 10). We further found publications relevant to the category Holistic view, which was introduced based on our pilot study. Publications directly relevant to all these 8 value categories were found in the main tracks of FSE and ICSE (Figure 10). Also, publications directly relevant to Security were found in all SE venues. Moreover, publications that directly considered *Benevolence* and *Universalism* were found across most venues/tracks. Publications directly relevant to

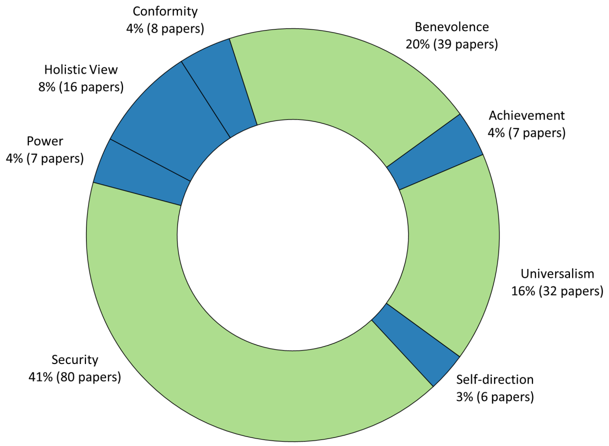

Figure 6: Directly Relevant publications per value category

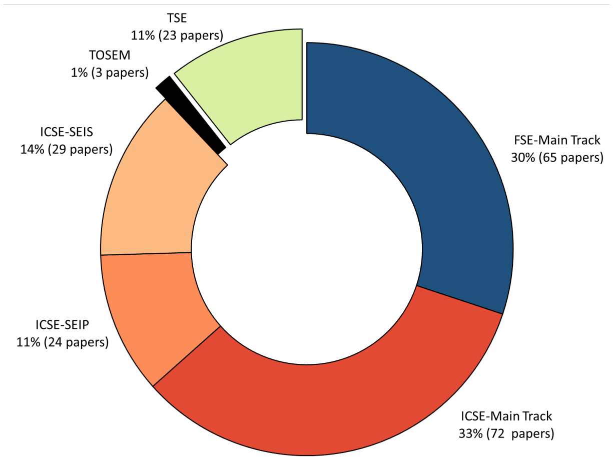

Figure 8: Directly relevant publications per venue/track

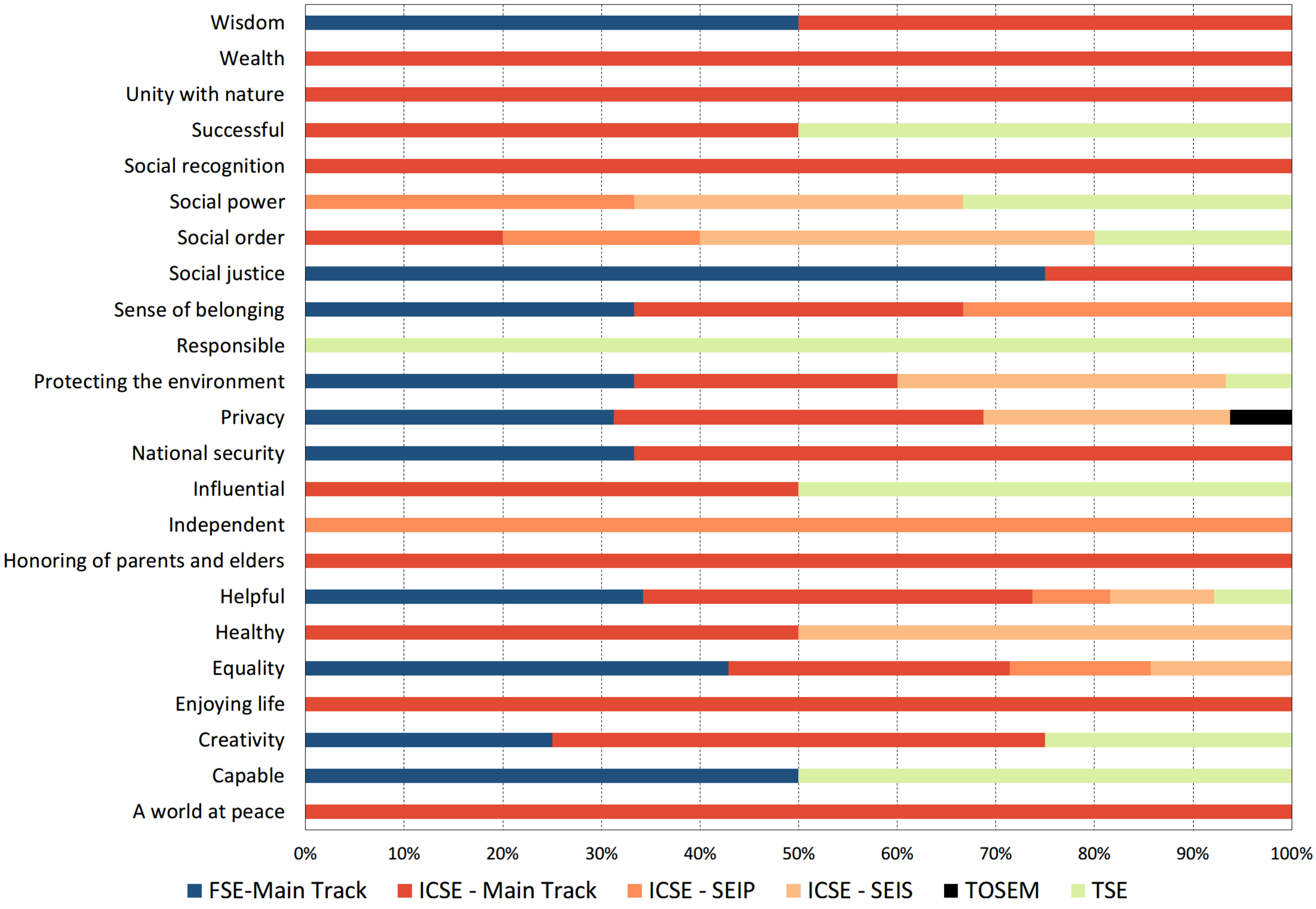

Figure 9: The distribution of directly relevant publications by venue/track: directly relevant publications were found only for 23 out of 58 Schwartz values (Figure 1)

Universalism were more prevalent in the SEIS track of ICSE. Publications in TOSEM only considered Security but not other value categories. It was also interesting to see that, compared to other venues/tracks, the SEIS track of ICSE contained the highest proportion of publications relevant to Conformity.

5 DISCUSSION

Our results indicate quite strongly that top SE research conferences and journals pay only limited attention to human values in software. Furthermore, of those papers classified as considering human values (16%), a significant proportion (41%) related to *Security*, thus implying that even where consideration of human values exists, it often tends to be about security issues.

It would be premature to conclude that SE research ignores human values. It may be that work on human values in SE exists in other SE conferences and journals, or indeed in other disciplinary areas, such as HCI or IS. Nevertheless, we argue that the lack of human values in leading general SE venues is problematic as it suggests either that SE researchers are not paying sufficient attention to the importance of human values, or that they are, but such work is not appearing in the leading SE venues, and hence, arguably not receiving the most visibility.

There are, however, two further considerations which affect how our results should be interpreted.

Firstly, one should not expect all SE papers to be directly relevant to values. For example, a paper describing a new static analysis technique is concerned chiefly with advancing the state of the art in static analysis not with broader questions of human values. Indeed, one might argue that most SE papers are of this ilk, and hence one should only aim for a relatively low percentage of values-relevant papers. Whilst valid, this argument begs the question as to what is the “target” percentage of values papers that the community should aim for? This is an open question, but we would argue that it should be higher than the current number because, as demonstrated in Section 4, 60% of human values were not considered at all. This means that 60% of values generally deemed to be important in society are ignored in leading SE research. Furthermore, if it is indeed the case that the reason for a low percentage of values papers is because most SE papers are deeply technical ones where broader human values are irrelevant, then this could be seen as

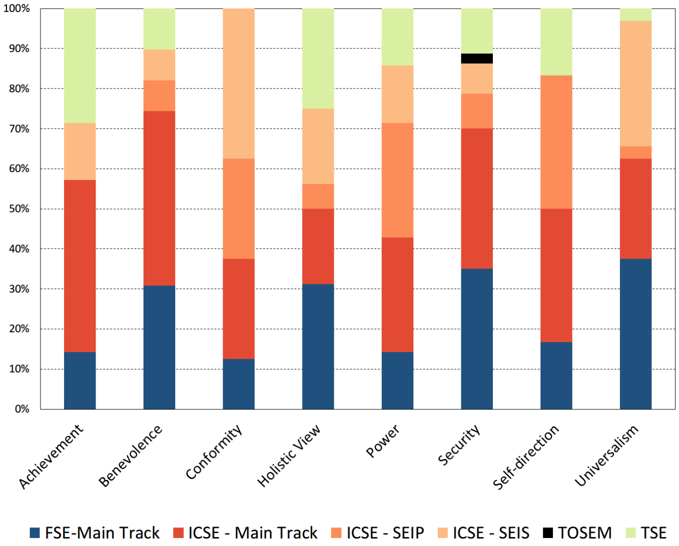

Figure 10: Publications directly relevant to different value categories across SE venues/tracks

a poor reflection of the community – should the community not strive to be more societally relevant?

Secondly, it can be argued that some of the Schwartz values are simply not relevant to software engineering. For example, it has often been commented that values like Mature Love or Obedience are outside the scope of SE. We argue the contrary. Software pervades every aspect of our society and increasingly, with the advent of artificial intelligence, makes decisions on our behalf. Surely, then, all values are relevant to software. Even values that may seem at first to be far from traditional SE, such as Mature Love, are relevant if we are to design software systems that promote or support love. If indeed that is the goal, then it needs to be captured in software requirements, designed for, and tested.

The Schwartz Value Structure (SVS) is just one model of human values. It was used in our research as it is by far the most widely adopted framework across a number of disciplines. It is a natural question to ask, however, whether the Schwartz model is the most appropriate in a SE context. There are two reasons to investigate whether Schwartz could be adapted to SE. Firstly, some of the nomenclature used in Schwartz is either different to that used by software engineering researchers or is unfamiliar to them. Mature Love, for example, could perhaps be rephrased to speak more easily to software engineers (cf. Google's "Don't be evil"). Or, we found in our study that many SE papers now talk about Sustainability, which does not appear directly in the Schwartz model, but must instead be captured as *Protecting the Environment* or *Unity with Nature.* Secondly, we found that the value category/value hierarchy did not always fit well with SE notions: for example, software engineers typically think of privacy as coming under a banner of security, whereas in Schwartz, *Privacy* comes under *Self-Direction.* As future work, therefore, it could be interesting to look at adapting and/or refining the Schwartz model to SE. Note that as our raters were not constrained by the Schwartz hierarchy, any concerns about the hierarchy's relevance to an SE context do not affect our overall results. Note also that extreme care needs to be taken if attempting to adapt Schwartz for SE. The Schwartz model has been validated for decades by many researchers across 20 different cultures; it is far from a non-trivial task to create a new values theory.

To summarize, we believe that our results indicate that the SE community might want to think about broadening its focus. This is not meant to be a judgement on the community but a spur towards considering more the social aspects of SE, in the same way that other areas of computer science, such as machine learning, are adapting their focus.

6 THREATS TO VALIDITY

In this section we discuss limitations of this research categorized as *Internal*, *External* and *Construct* validity threats.

Construct Validity: choosing a classification scheme suited for the software engineering domain was one of the main challenges for this research. In the absence of an SE-specific scheme to classify human values, we selected SVS, a well established theory from social sciences to study and explain human values [44] that has been successfully applied in SE [13, 26, 48]. SVS was adopted in this research as an independent classification scheme, instead of developing our own, to mitigate the risk of introducing researcher bias. The definition of 'directly relevant' was crucial to the classification of selected publications. The definition was therefore, carefully developed as a criteria to allow classification of research mainly focused on addressing human values but at the same time avoid classifying almost every paper as 'relevant' merely on the premise that 'all research is helpful or useful'.

Similar to Glass et al. [18], lack of mutual exclusion was a challenge for our classification scheme. It was often possible to classify a paper as relating to more than one individual value. This we believe was more to do with the ill-defined nature of human values than a limitation of the chosen classification scheme. Still, the potential threat was mitigated by using an iterative process and conducting rater training to better understand and clarify relationships between values and their categories.

In some cases, the raters found that certain papers were related to human values in general rather than to any particular value. Forcing such papers into a single value category would have influenced results. To mitigate this, we added a new *Holistic View* category to our classification scheme.

Internal Validity threats for this study arise from the complexities of categorizing papers into the selected classification scheme. It is possible that the raters' own expertise in understanding the scheme categories and definitions of values may have influenced paper classifications. This risk was mitigated as the classification process forced random assignment of each paper to two raters and in case of a disagreement an independent arbiter was introduced to facilitate agreement. Some disagreements (2%, see Figure 2) remained even after the arbiter's intervention. In such cases we did not force consensus.

While a detailed review of the entire papers (rather than just the abstract, title and keywords) could have provided more accurate results, we adopted a procedure similar to those used in previous studies [6, 37] published in a top SE conference (ICSE) and a respected SE journal (JSS).

External validity threats may arise from potential limitations of our choice of publication venues and the block of time period under study (i.e., 2015-2018). The chosen venues are widely acknowledged as the top-tier venues of SE research; however, we accept that the results may be different if other more specialist conferences/journals had been considered.

Generalizability of results based on a subset of papers is often a concern for empirical studies. In our research, this risk was mitigated by using 1350 papers published in the last 4 years which can be considered a good representation of trends in SE research as suggested in [6]. The findings of this study, however, may be biased towards ICSE and FSE as they published more papers in the selected period compared to journals (ICSE 559, FSE 512 vs. TSE 215 and TOSEM 64).

7 RELATED WORK

Classification of papers has been widely adopted in the SE literature [23, 37, 42, 46, 48] as a way of providing insights on trends and directions in SE research. Such findings, though not conclusive, can indicate the general attitude of SE researchers as well as the priorities in SE research. Paper classification helps to highlight the gaps and the needs for further research in specific SE domains. Mary Shaw [37], for instance, analyzed the abstracts of research papers submitted and accepted to ICSE 2002 to identify different research types, trends in research questions, contribution types and validation approaches. The author also studied the program committee discussions regarding the acceptance or rejection of the papers. Another example is the work by Vessey et al. [46] who categorized samples of SE papers published from 1995 to 1999 in six journals based on topic, method, and approach. Another study by Williams et al. [48] classified ICSE publications from 2015-2017 using a framework developed in psychology and sociology as a lens to understand how SE research captures human and social perspectives.

However, paper classification methods rely on classification schemes, that can be general or specific depending on the purpose of the classification. To classify different SE papers, Montesi and Lago [23] presented a classification approach based on the call for papers of top-tier SE conferences and journals included in the Journal Citation Reports and the instructions to authors of relevant journals and published works. Also, Ioannidis et al. [21] categorized the meta-research discipline into five main thematic fields corresponding to how to conduct, report, verify, correct and reward science. There have also been efforts to develop specific classification schemes. For instance, Wieringa et al. [47] developed a classification scheme to identify papers that belong to Requirements Engineering as a subdomain in SE. Sjoberg et al. [38] surveyed SE papers in nine journals and three conferences from 1993 to 2002 with the aim to characterize controlled experiments in SE by characterizing the topics of the experiments and their subjects, tasks, and environments.

Moreover, some paper classifications have identified gaps in SE practice. An example is the work by Stol and Fitzgerald [41], where the authors observed the lack of a holistic view in SE research. The work contributed a framework for positioning a holistic set of research strategies and showed its strengths and weaknesses in relation to various research components. Also, Zelkowitz and Wallace [51] classified, according to a 12-model classification scheme, around 600 SE papers published over a period of three years to provide insights on the use of experimentation within SE. They identified a gap in SE research with respect to validation and experimentation. Another example is an empirical study of SE papers performed by Zannier et al. [50] to investigate the improvement of the quantity and quality of empirical evaluations conducted within ICSE papers over time. The authors compared a random sample of papers in two periods, 1975 − 1990 and 1991 − 2005, and found that the quantity of empirical evaluation has grown, but the soundness of evaluation has not grown at the same pace.

Last but not the least, some paper classifications have provided insights on SE venues in relation to the papers published in those venues. An example is the work by Systa et al. [42] that investigated the turnover of PC compositions and paper publication in six SE conferences. The work was later extended by Vasilescu et al. [45] by proposing a wider collection of metrics to assess the health of 11 SE conferences over a period of more than 10 years.

8 CONCLUSIONS AND FUTURE WORK

Repeated incidents of software security and privacy violations continue to attract researchers' attention. In this paper, however, we investigated the prevalence of a broader range of human values including *Trust*, *Equality* and *Social Justice* in software engineering research. Using the Schwartz Values Structure as our classification scheme, which identifies 58 human values, we classified 1350 papers recently published (2015–2018) in top-tier SE conferences and journals. We conclude that only a small proportion of SE research in leading venues directly considers human values. While *Security* and *Privacy* stand out as the main focus in SE research, a broad range of human values remain inadequately addressed in leading SE venues. Finally, we found that leading SE conferences publish more values relevant research compared to leading SE journals.

In the future, we want to extend this study using a machine learning approach. Manually labelled data from this study will be used for training machine learning algorithms to classify larger sets of publications with the aim to better visualize how SE addresses human values. We also plan to utilise our manually labelled data captured from various SE contexts to develop definitions of human values that are relatively easy for practitioners to understand and implement. Finally, we plan to carry out case studies in software organizations to investigate whether SE research related to human values has actually made an impact on SE practice.

REFERENCES

[1] Huib Aldewereld, Virginia Dignum, and Yao-hua Tan. 2015. Design for values in software development. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (2015), 831–845.

[2] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] M. Bagherzadeh, N. Kahani, C. Bezemer, A. E. Hassan, J. Dingel, and J. R. Cordy. 2018. [Journal First] Analyzing a Decade of Linux System Calls. In *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. 267–267. https://doi.org/10.1145/3180155.3182518

[4] Nick Baker. 2019. Molly Russell: Instagram bans graphic self-harm images after suicide of UK teen. https://www.sbs.com.au/news/molly-russell-instagram-bans-graphic-self-harm-images-after-suicide-of-uk-teen

[5] Juan A Barceló, Florencia Del Castillo Bernal, Ricardo Del Olmo, Laura Mameli, FJ Miguel Quesada, David Poza, and Xavier Vilà. 2014. Social interaction in hunter-gatherer societies: simulating the consequences of cooperation and social aggregation. *Social Science Computer Review* 32, 3 (2014), 417–436.

[6] Antonia Bertolino, Antonello Calabrò, Francesca Lonetti, Eda Marchetti, and Breno Miranda. 2018. A categorization scheme for software engineering conference papers and its application. *Journal of Systems and Software* 137 (2018), 114–129. https://doi.org/10.1016/j.jss.2017.11.048

[7] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics* 24, 2 (2018), 505–528.

[8] An-Shou Cheng and Kenneth R Fleischmann. 2010. Developing a meta-inventory of human values. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47*. American Society for Information Science, 3.

[9] Paresh Dave. 2018. Google bars uses of its artificial intelligence tech in weapons. https://www.reuters.com/article/us-alphabet-ai/google-bars-uses-of-its-artificial-intelligence-tech-in-weapons-idUSKCN1J32M7.

[10] Roel van Dijk, Christophe Creeten, Jeroen van der Ham, and Jeroen van den Bos. 2017. Model-driven software engineering in practice: privacy-enhanced filtering of network traffic. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 860–865.

[11] Amitai Etzioni and Oren Etzioni. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21, 4 (2017), 403–418.

[12] Robert Feldt. 2016. ISI SE Journals (Ranked) http://www.robertfeldt.net/advice/se_venues/.

[13] Maria Angela Ferrario, Will Simm, Stephen Forshaw, Adrian Gradinar, Marcia Tavares Smith, and Ian Smith. 2016. Values-first SE: research principles in practice. In *Proceedings of the 38th International Conference on Software Engineering Companion*. ACM, 553–562.

[14] Maria Angela Ferrario, Will Simm, Peter Newman, Stephen Forshaw, and Jon Whittle. 2014. Software Engineering for 'Social Good': Integrating Action Research, Participatory Design, and Agile Development. In *Companion Proceedings of the 36th International Conference on Software Engineering (ICSE Companion 2014)*. ACM, New York, NY, USA, 520–523. https://doi.org/10.1145/2591062.2591121

[15] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. 2005. Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 751–760.

[16] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[17] Batya Friedman and Peter H Kahn Jr. 2007. Human values, ethics, and design. In *The human-computer interaction handbook*. CRC Press, 1223–1248.

[18] Robert L. Glass, Iris Vessey, and Venkataraman Ramesh. 2002. Research in software engineering: an analysis of the literature. *Information and Software technology* 44, 8 (2002), 491–506.

[19] Preston Gralla. 2016. Amazon Prime and the racist algorithms. https://www.computerworld.com.au/article/599661/amazon-prime-racist-algorithms.

[20] Tim Holmes, Elena Blackmore, Richard Hawkins, and Tom Wakeford. 2011. *The common cause handbook*. Public Interest Research Center.

[21] John PA Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N Goodman. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLoS biology* 13, 10 (2015), e1002264.

[22] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[23] Michela Montesi and Patricia Lago. 2008. Software engineering article types: An analysis of the literature. *Journal of Systems and Software* 81, 10 (2008), 1694–1714.

[24] Rodrigo Morales, Rubén Saborido, Foutse Khomh, Francisco Chicano, and Giuliano Antoniol. 2018. Earmo: An energy-aware refactoring approach for mobile apps. *IEEE Transactions on Software Engineering* 44, 12 (2018), 1176–1206.

[25] Davoud Mougouei, Harsha Perera, Waqar Hussain, Rifat Shams, and Jon Whittle. 2018. Operationalizing human values in software: a research roadmap. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018*. 780–784. https://doi.org/10.1145/3236024.3264843

[26] Md Saddam Hossain Mukta, Euna Mehnaz Khan, Mohammed Eunus Ali, and Jalal Mahmud. 2017. Predicting movie genre preferences from personality and values of social media users. In *Eleventh International AAAI Conference on Web and Social Media*.

[27] Mark O Riedl and Brent Harrison. 2016. Using stories to teach human values to artificial agents. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[28] Milton Rokeach. 1973. *The nature of human values*. Free press.

[29] Justin Sablich. 2017. 'Price Gouging' and Hurricane Irma: What Happened and What to Do. https://www.nytimes.com/2017/09/17/travel/price-gouging-hurricane-irma-airlines.html.

[30] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Vol. 25. Elsevier, 1–65.

[31] Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues* 50, 4 (1994), 19–45.

[32] Shalom H Schwartz. 2005. Basic human values: Their content and structure across countries. *Valores e comportamento nas organizações* (2005), 21–55.

[33] Shalom H Schwartz. 2006. Les valeurs de base de la personne: théorie, mesures et applications. *Revue française de sociologie* 47, 4 (2006), 929–968.

[34] Shalom H Schwartz. 2007. Basic human values: Theory, methods, and application. *Risorsa Uomo* (2007).

[35] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2, 1 (2012), 12–13. https://doi.org/10.9707/2307-0919.1116

[36] Shalom H Schwartz and Klaus Boehnke. 2004. Evaluating the structure of human values with confirmatory factor analysis. *Journal of research in personality* 38, 3 (2004), 230–255.

[37] Mary Shaw. 2003. Writing good software engineering research papers. In *Software Engineering, 2003. Proceedings. 25th International Conference on*. IEEE, 726–736.

[38] Dag IK Sjøberg, Jo Erskine Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, N-K Liborg, and Anette C Rekdal. 2005. A survey of controlled experiments in software engineering. *IEEE transactions on software engineering* 31, 9 (2005), 733–753.

[39] David Smith. 2018. Zuckerberg put on back foot as House grills Facebook CEO over user tracking https://www.theguardian.com/technology/2018/apr/11/zuckerberg-hearing-facebook-tracking-questions-house-back-foot.

[40] Igor Steinmacher, Tayana Uchoa Conte, Christoph Treude, and Marco Aurélio Gerosa. 2016. Overcoming Open Source Project Entry Barriers with a Portal for Newcomers. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. ACM, New York, NY, USA, 273–284. https://doi.org/10.1145/2884781.2884806

[41] Klaas-Jan Stol and Brian Fitzgerald. 2015. A holistic overview of software engineering research strategies. In *Proceedings of the Third International Workshop on Conducting Empirical Studies in Industry*. IEEE Press, 47–54.

[42] Tarja Systä, Maarit Harsu, and Kai Koskimies. 2012. Inbreeding in Software Engineering Conferences.

[43] Christopher Theisen, Marcel Dunaiski, Laurie Williams, and Willem Visser. 2017. Writing good software engineering research papers: revisited. In *Proceedings of the 39th International Conference on Software Engineering Companion*. IEEE Press, 402–402.

[44] Sarah Thew and Alistair Sutcliffe. 2018. Value-based requirements engineering: method and experience. *Requirements Engineering* 23, 4 (2018), 443–464.

[45] Bogdan Vasilescu, Alexander Serebrenik, Tom Mens, Mark GJ van den Brand, and Ekaterina Pek. 2014. How healthy are software engineering conferences? *Science of Computer Programming* 89 (2014), 251–272.

[46] Iris Vessey, Venkataraman Ramesh, and Robert L Glass. 2002. Research in information systems: An empirical study of diversity in the discipline and its journals. *Journal of Management Information Systems* 19, 2 (2002), 129–174.

[47] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. 2006. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requirements Engineering* 11, 1 (2006), 102–107.

[48] Courtney Williams, Margaret-Anne Storey, Neil A. Ernst, Alexey Zagalsky, and Eirini Kalliamvakou. 2019. Methodology Matters: How We Study Socio-Technical Aspects in Software Engineering. *ACM Transactions on Software Engineering* 37, 4 (2019). arXiv:1905.12841 http://arxiv.org/abs/1905.12841

[49] Emily Winter, Steve Forshaw, and Maria Angela Ferrario. 2018. Measuring human values in software engineering. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 48.

[50] Carmen Zannier, Grigori Melnik, and Frank Maurer. 2006. On the success of empirical studies in the international conference on software engineering. In *Proceedings of the 28th International Conference on Software Engineering*. ACM, 341–350.

[51] Marvin V Zelkowitz and Dolores Wallace. 1997. Experimental validation in software engineering. *Information and Software Technology* 39, 11 (1997), 735–743.