SAFE-PASS: Stewardship, Advocacy, Fairness and Empowerment in Privacy, Accountability, Security, and Safety for Vulnerable Groups

Indrajit Ray
Colorado State University
Fort Collins, CO, U.S.A.
indrajit.ray@colostate.edu

Bhavani Thuraisingham
University of Texas at Dallas
Dallas, TX, U.S.A.
bhavani.thuraisingham@utdallas.edu

Jaideep Vaidya
Rutgers University
Newark, NJ, U.S.A.
jsvaidya@business.rutgers.edu

Sharad Mehrotra
University of California at Irvine
Irvine, CA, U.S.A.
sharad@ics.uci.edu

Vijayalakshmi Atluri
Rutgers University
Newark, NJ, U.S.A.
atluri@rutgers.edu

Indrakshi Ray
Colorado State University
Fort Collins, CO, U.S.A.
indrakshi.ray@colostate.edu

Murat Kantarcioglu
University of Texas at Dallas
Dallas, TX, U.S.A.
muratk@utdallas.edu

Ramesh Raskar
Massachusetts Institute of Technology
Cambridge, MA, U.S.A.
raskar@media.mit.edu

Babak Salimi
University of California at San Diego
San Diego, U.S.A.
bsalimi@ucsd.edu

Steve Simske
Colorado State University
Fort Collins, CO, U.S.A.
steve.simske@colostate.edu

Nalini Venkatasubramanian
University of California at Irvine
Irvine, CA, U.S.A.
nalini@uci.edu

Vivek Singh
Rutgers University
New Brunswick, NJ, U.S.A.
v.singh@rutgers.edu

ABSTRACT

Our vision is to achieve societally responsible secure and trustworthy cyberspace that puts algorithmic and technological checks and balances on the indiscriminate sharing and analysis of data. We achieve this vision in a holistic manner by framing research directions with four major considerations: (i) Expanding knowledge and understanding of security and privacy perceptions and expectations in vulnerable groups, which significantly contribute to their unwillingness to share data, and use that knowledge to drive research in (a) mitigating missing/imbalanced data problems, (b) understanding and modeling security and privacy risks of data sharing, and (c) modeling utility of data sharing. (ii) Developing a risk-adaptive, policy model capable of capturing and articulating security and privacy expectations of users that are relevant in a particular context and develops associated technology to ensure provenance and accountability. (iii) Developing robust AI/ML algorithms that are transparent and explainable with respect to fairness and bias to reduce/eliminate discrimination, misuse, privacy violations, or other cyber-crimes. (iv) Developing models and techniques for a nuanced, contextually adaptive, and graded privacy paradigm that allows trade-offs between privacy and utility. Towards this, in this paper we present the SAFE-PASS framework to provide Stewardship, Advocacy, Fairness and Empowerment in Privacy, Accountability, Security, and Safety for Vulnerable Groups.

SACMAT '23, June 7–9, 2023, Trento, Italy.
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0173-3/23/06.
https://doi.org/10.1145/3589608.3593830

CCS CONCEPTS

• **Security and privacy → Privacy-preserving protocols**; **Authorization**; **Access control**; **Pseudonymity, anonymity and untraceability**; **Social aspects of security and privacy**; **Privacy protections**; **Usability in security and privacy**.

KEYWORDS

privacy, security, usability, accountability, fairness, machine learning, vulnerable populations

ACM Reference Format:
Indrajit Ray, Bhavani Thuraisingham, Jaideep Vaidya, Sharad Mehrotra, Vijayalakshmi Atluri, Indrakshi Ray, Murat Kantarcioglu, Ramesh Raskar, Babak Salimi, Steve Simske, Nalini Venkatasubramanian, and Vivek Singh. 2023. SAFE-PASS: Stewardship, Advocacy, Fairness and Empowerment in Privacy, Accountability, Security, and Safety for Vulnerable Groups. In *Proceedings of the 28th ACM Symposium on Access Control Models and Technologies (SACMAT '23), June 7–9, 2023, Trento, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3589608.3593830

1 INTRODUCTION

The flood of data available via tracking users' activities online and in social media, via surveillance and smart sensing and the advances made in Big Data, AI, and ML for classifying, interpreting, and analyzing this data, hold tremendous potential for bettering the quality of life and health of the world. Unfortunately, the indiscriminate use of the data and associated technologies disproportionately and negatively impact vulnerable groups via exploitation of inherent biases in data and algorithms, misclassification, opportunities for malicious misinterpretation and misuse, leading to further privacy leakage. Cybercrimes centered around compromised / comprising information disproportionately affect these groups, since they are often motivated to maintain a low profile for a variety of reasons. They are also very much dependent on surrogates (or proxies) to help carry out many of their day-to-day activities. This indirection translates to larger opportunities for data breaches and misuses. Unfortunately, often there is a translation/interpretation gap of security and privacy requirement specification, services delivered, and/or alerts and messages when conveyed through the proxies, resulting in poorer inculcation of trust in technology for vulnerable groups.

The goal of this work is to explore socio-technical approaches to support *privacy, accountability, security, and safety* (PASS) that can help create the next generation of responsible information technology systems designed to make a positive difference for the vulnerable population, while providing *stewardship, advocacy, fairness, and empowerment* (SAFE). Towards this goal, we propose a new vision of societally responsible security and privacy called SAFE-PASS. The research towards achieving SAFE-PASS is inter-disciplinary in nature requiring advances in security, privacy and trust, risk modeling, machine learning, algorithmic fairness, computational social science, data analytics and management, software engineering, formal methods system design, data, geriatrics, health and well-being, and societal systems suitable for the vulnerable population.

The organization of this paper is as follows: We identify some defining characteristics of vulnerable groups in Section 2 and discuss how we choose two representative groups to study in the SAFE-PASS context. Section 3 discusses the SAFE-PASS vision towards achieving privacy, accountability, security and safety for our vulnerable groups. This is followed by a discussion on the research that we are undertaking towards achieving this vision in Section 4. In particular, Subsection 4.1 gives an overview of our research in adaptive policies and accountability, Subsection 4.2 discusses our approach to utility-driven adaptive policies, Subsection 4.3 presents the research in AI/ML algorithmic fairness and bias and Subsection 4.4 describes the SAFE-PASS technology realization architecture. We conclude with some parting thoughts in Section 5.

2 CHARACTERISTICS OF VULNERABLE GROUPS AND THEIR SECURITY AND PRIVACY CHALLENGES

Vulnerable Groups: While reference to vulnerable populations in medical healthcare, ethics, and legal research abound and several protective guidelines exist to offer special protections of the vulnerable population (see for example, [4]), the concept of vulnerability and hence the criteria using which a particular population can be deemed as vulnerable remains ambiguous. Often, vulnerable populations are identified explicitly as children, prisoners, pregnant women, handicapped, mentally disabled, economically disadvantaged, or educationally deficient. To support our perspective of designing secure data-driven approaches, we consolidate our observations from the literature into characterizing vulnerable groups as those that meet one or more of the following criteria: (i) Inability to make informed decisions and hence requiring proxy/surrogate with appropriate power of attorney. (ii) Susceptible to being unduly influenced by others to a degree that might be detrimental to their well-being. (iii) Limited capability for self-protective actions from intended or inherent risk and consequently dependent upon others for their well-being. (iv) Limited in their freedom to act, speak, or think as they want without hindrances or restraints. Have no control over or awareness of how their data is used (that can lead to discrimination, unfair treatment, and data monetization). (v) May experience intense fear for their safety because of earlier experiences in life. In addition, we characterize vulnerable populations into two categories: (a) Those who are explicitly vulnerable, i.e., whose status as being part of a vulnerable group itself is not sensitive (e.g., elderly living in assisted living facilities), and (b) Those whose vulnerable status is itself sensitive and must be hidden to appropriately protect them (e.g., victims of human trafficking, or those who might not have a legal status to stay in the country).

While SAFE-PASS research on building responsible IT systems of the future will transcend across diverse vulnerable groups, to bring focus to our research and to leverage our ongoing collaborations, we work closely with two representative groups – (a) Elderly population living alone or a in shared/assisted-living facilities, and (b) Victims of human trafficking. The former belongs to the explicitly vulnerable group exemplifying the criteria (i)-(iv) above while the latter belongs to the hidden vulnerable group exemplifying the criteria (ii)-(v).

Information Technology and its Evolution The past two decades have witnessed unparalleled advances in technology: People connected to each other, network in the hands of everyone, and knowledge and services on demand. More recent advances in sensing and actuation and ambient intelligence via big data analytics (BD), artificial intelligence (AI) and machine learning (ML) hold enormous potential to benefit these vulnerable groups. At the same time, the *untrammelled and unrestricted* use of the same technology, coupled with issues related to flaws, and algorithmic biases and fairness, can equally cause enormous harm to the same groups. The following two scenarios illustrate this dichotomy between the positive and negative implications of technology.

Example 1: Alice, elderly lady, living alone: Smart technology currently exists that can monitor Alice discreetly or sense whether she has fallen. Imagine the near future. A smart robot assistant is the 24/7 companion of Alice. The smart robot assistant holds the hand of the woman and assists her while she moves around her room. It guides her to the balcony to watch the sunrise and the sunset. It opens the door to let Alice's grandchildren in when they visit her. The assistant keeps in safe custody her living will, various powers of attorney, medicine schedule, and even her health records. One day, Alice feels dizzy and faints. The smart assistant summons the medics, lets them in. The medics in turn coordinate her care with a remote doctor who administers a pill of smart medical nano-bots for Alice to take. The nano-bots provide a detailed image of the aneurysm in her brain to the remote doctor and allow the doctor to precisely perform a surgery to fix the problem.

Data sharing and misuse causing human rights violation: On the urging of her granddaughter who is working on a class project, Alice sends in her DNA sample to 23AndUs taking advantage of a company promotion offering free genetic testing for family tree. Unfortunately, she could not read the (really) fine print. She thinks the “23” refers to the number of chromosomes in the human karyotype. Little did she suspect that the “23” meant the number of interested data aggregators to whom 23AndUs has passed along her DNA and its immutable biometric. Annoying at first, but then the unthinkable happens, and there is a global pandemic whose individual severity is dependent on the extent and location of operons that modulate the expression of both intron and exon (intervening and expressed) sequences in the DNA. The data, long since shared with the partners of 23AndUs, is now mined to find individual susceptibility to a yet undiscovered variant of COVID-22. Finally, the mined DNA finds itself on the web from where the neighborhood restaurant comes to know about Alice's susceptibility. The restaurant owner refuses to let Alice be seated for dinner.

Example 2: Jane, victim of human trafficking, rehabilitating in her home: Jane is extremely fearful for her physical safety and a breach of her confidential information can mean life or death for her. Jane has installed Ringo 3.0, a new biometric (face and eyes) recognition-based security system, for her house. Ringo 3.0 is also a safety net for her; it can take Jane's picture, selectively remove features to protect her sensitive information and then analyze it to understand her specific situation and provide necessary support. Jane interacts with the outside world via a case worker who visits her regularly to help in her therapy and recovery. Then COVID-19 strikes, and the case worker is no longer reachable. Ringo 3.0 connects Jane with a chat-bot case worker which steps in place of the human case worker. The chat bot pulls up information from Jane's security system and guides, advises, and comforts her to relieve her anxiety. One day Jane believes she heard the voice of one of her tormentors and begins to hyperventilate. The security system alerts the chat bot which summons an autonomous vehicle to take Jane to safety.

Algorithmic bias and fairness problem leading to discrimination and significant mental trauma: Ringo 3.0 has been a big safety net for Jane. However, Ringo 3.0 was updated with a new ML model based on a broader training set. This results in false negative errors for those least represented in the training data. Jane now finds herself locked out of her own house and left explaining herself to the police officers who arrive when alerted by Ringo 3.0. One day Jane starts to become fearful again. Ringo 3.0 assesses the situation. However, Jane is paranoid. Her demeanor and language used in soliloquy lead Ringo 3.0's algorithm to believe that there is a hidden intruder in the house who is involved in human trafficking. Ringo 3.0. directly contacts the police. A new police officer comes, and arrests Jane based on information shared by Ringo 3.0.

Why are these vulnerable groups most impacted? The marginalized, economically disadvantaged, and those in the lowest quintile of income are most affected by these breaches. This is because, as shown in Figure 1, the *use/misuse of poorly designed technology* and *indiscriminate use of data lead* to problems with privacy, accountability, security, and safety, negatively impacting the well-being of vulnerable groups. Their reliance on surrogates who often have more access to their sensitive information makes them more susceptible to persuasion, extortion and other crimes of compromised data and misinformation.

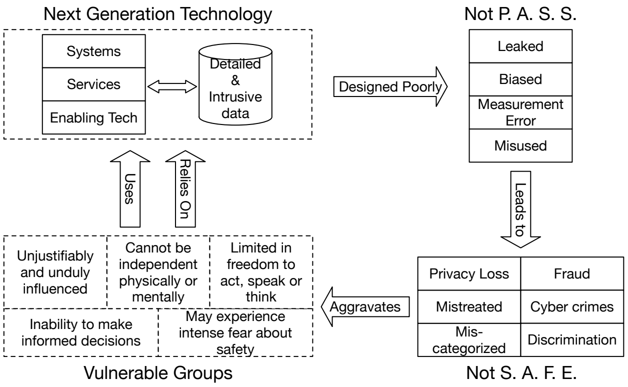
Next Generation Technology
Systems
Services
Enabling Tech
Detailed & Intrusive data
Designed Poorly
Not P. A. S. S.
Leaked
Biased
Measurement Error
Misused
Uses
Relies On
Leads to
Unjustifiably and unduly influenced
Cannot be independent physically or mentally
Limited in freedom to act, speak or think
Inability to make informed decisions
May experience intense fear about safety
Vulnerable Groups
Aggravates
Privacy Loss
Fraud
Mistreated
Cyber crimes
Mis-categorized
Discrimination
Not S. A. F. E.

Figure 1: Data flow across poorly designed technology results in security, privacy and fairness issues for vulnerable groups

Vulnerable groups, especially the hidden vulnerable, are often not part of the training population of AI/ML models. They do not know who has access to their data and/or how it is used and so are fearful about sharing data. Data imbalance and missing data result in benefits skewed in favor of the privileged elements of society. Instances of discrimination via mis-categorization and misidentification (such as in obtaining health insurance, in spread of misinformation, in lending or job searches) that can be attributed to biases in the data and/or the techniques are increasing. These problems often arise from measurement error, misclassification, and selection bias in the data, in addition to adversarial attacks on the data. Biases can also be introduced by the handling and transformation of data throughout the ML processing pipeline including model development and deployment. There has been an increase in cyber-crimes, such as phishing and social engineering attacks, cyber-extortion, cyber-bullying, and cyber-stalking, resulting from leaked and compromised private data and from manipulated and misconstrued data. The latter can often be directly attributed to different biases that result in mis-categorization.

Resources available to more privileged group to protect themselves against these problems are much less readily accessible to vulnerable groups. There are also fewer and fewer opportunities of inculcating trust in technology. There is frequently a translation / interpretation gap of security and privacy requirement specification, services delivered and/or alerts and messages when conveyed through the surrogates. For instance, the vulnerable people may want to take shortcuts OR at least not want to enable aspects like two factor authentication to “keep things simple” and not ““unnecessarily trouble” their surrogates. While data sharing, data analysis etc. have immense value, the benefit value must be explained to inculcate trust in the system.

3 THE SAFE-PASS VISION

What is SAFE-PASS? We envision a fundamentally transformed society where security and privacy (S&P) techniques are designed and used in a trustworthy and explainable manner, where privacy is not an all-or-nothing property but is nuanced based on the value of revealing information to the data owner, where vulnerable groups are not victimized by S&P and associated BD/AI/ML technology solely because they fail to understand how the technology works or understand if it is working against their best interests, and begin to appreciate where S&P comes into play to help them.

SAFE-PASS is envisioned around a new societally responsible data integration, analysis and sharing paradigm for S&P, termed *Selective Secrecy and Structural Transparency*. This is a *computationally reflective* model [33, 35]. In a reflective architecture, a self-observing, introspecting meta-system continuously observes the state of the underlying system/platform at various granularities, analyze the collected observations and enable the system to adapt to perturbations and changes associated with information flow. The resulting execution model follows an "Observe - Analyze - Adapt" (O-A-A) loop.

Under this new S&P paradigm, Selective Secrecy involves judiciously providing strong levels of security and privacy to shared data by default and updating the levels based on situational awareness and an evaluation of the utility of the sharing. It requires evaluating the trade-offs between privacy and security risks and the utility of sharing keeping the expectations of vulnerable groups in mind. The expectations are driven by fairness and explainability. It requires newer ways of policy specification, design, analysis, and provable enforcement. Structural Transparency involves enabling questions such as, "what information about me is out there," "am I being misidentified or miscategorized," "is my information being used against me," or "is the data requested too invasive," to be asked and answered and preventive actions to be taken. It requires a fundamental shift in how data is collected, extracted, aggregated, and shared. It requires careful consideration of data integration, feature selection, modeling development and deployment. It requires new foundational results in provenance for ML models for operationalizing transparency, accountability, auditability and debugability.

Through the proposed research in SAFE-PASS, we aim to achieve stewardship, advocacy, fairness, and empowerment of vulnerable groups in privacy, accountability security and safety. *Stewardship* involves developing technology for the selective secrecy and structural transparency of vulnerable groups and guiding them to make informed decisions. *Advocacy* involves proactively evaluating technology via working with partnering organizations to raise awareness, identify, develop, and adopt best practices, policies and technologies and disseminate knowledge. *Fairness* is achieved by developing tools and techniques that mitigate biases and augment data to emphasize vulnerable groups. Finally, *Empowerment* results from additional means of safe and private access to information via education and training. We are hopeful that SAFE-PASS would help to re-vitalize the social contract that our ancestors wrote for a representational democracy.

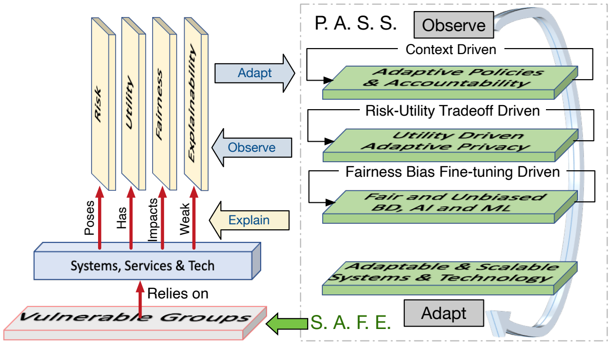
P. A. S. S.
Observe
Context Driven
Adaptive Policies & Accountability
Risk-Utility Tradeoff Driven
Utility Driven Adaptive Privacy
Fairness Bias Fine-tuning Driven
Fair and Unbiased BD, AI and ML
Adaptable & Scalable Systems & Technology
Adapt
S. A. F. E.
Risk
Utility
Fairness
Explainability
Adapt
Observe
Explain
Poses
Has
Impacts
Weak
Systems, Services & Tech
Relies on
Vulnerable Groups

Figure 2: The SAFE-PASS approach to achieving societally responsible S&P systems

4 RESEARCH PATHWAYS TOWARDS ACHIEVING SAFE-PASS

The SAFE-PASS approach to achieving selective secrecy and structural transparency is shown in Figure 2.

It involves research in four different but interlinked areas: (1) Adaptive Policies and Accountability, (2) Utility Driven and Usable Adaptive Policies, (3) Fair and Unbiased Big Data / AI / ML, and (4) Adaptable and Scalable Systems and Technology. The expectations of vulnerable groups in sharing data and the utility of sharing (benefits, perceived or not-perceived but explainable) drives the policies research in pathway (1). These policies drive our research in areas (2) and (3) to provide societally responsible means of data collection, extraction, aggregation and sharing. Finally, the algorithms and techniques are implemented in research (4). In the following, we give an overview of the research and explore challenges and next steps to bring the vision of SAFE-PASS to fruition.

4.1 Adaptive Policies and Accountability

Policies in systems may arise through regulations, such as GDPR [31], laws to protect the vulnerable population (see for example, [5, 6]), or in the form of preferences by the individuals (either through the vulnerable population, by themselves, or through their surrogates). The first challenge in supporting policies is to extract them from regulations and/or elicit them from individuals. Organizational policies including regulations such as GDPR are often expressed in natural languages which raises challenges of ambiguity in their interpretation. Another challenge arises from the semantically higher level of abstraction at which these policies are expressed to be understandable to humans. For instance, depending upon the vulnerable group, there could be sensitivity associated with leakage of the precise location and/or daily routine of an individual. Such a semantically higher-level observation about individual's location could be inferred through, often innocuous, lower-level data which, if accessible to the adversary, could lead to leakage. Our prior work has shown numerous situations wherein data captured through sensing devices (e.g., motion sensors, power sensors) could lead to leakages of social, cultural, religious, and health-related habits (e.g., smoker/non-smoker) [24].

SAFE-PASS goal is to protect sensitive data of vulnerable groups and provide fine-grained access to the various stakeholders where the security and privacy needs are *context dependent*. This access to data needs to be evaluated against potential risks to vulnerable population as well as the potential for abuse and attacks. With this in mind, we focus on several key areas:

Policy requirements elicitation: Security, privacy, and accountability policies are derived from artifacts expressed in natural languages. These include privacy and security concerns expressed in surveys, organizational rules, and state and federal regulations. Focus group interviews following a semi-structured approach[25, 30] can be conducted with multiple stakeholders including vulnerable group members, their personal caregivers, health/service professionals, and system designers to understand the privacy and security needs of the vulnerable populations and how they differ from those in the general population. The themes emerging from this analysis (e.g., via grounded theory) can help frame the design requirements.

Policy model generation: Manual inspection of requirements document reveals the components of interest from the policy formalization perspective. These include statements pertaining to policies, attributes of the various stakeholders, characteristics of the data, and the context of data sharing. Formal models of policies, that allow for controlled delegation and context representation are being developed. SAFE-PASS focuses on automated formal policy generation from natural language statements. Most works in this area use rules or linguistic features of the language [1, 38] and focus only on resource access control without considering obligation, context-based, or administrative policies. Instead, contextual pre-trained transformer-based language models [7, 32] can be used and combined with semantic role labeling units like [23, 26] that are used to discover the predicate-argument structure in a sentence.

Policy analysis and evolution: As we continue to generate appropriate policies from requirement specification we need to analyze the policies to determine if the policies have any conflict and if they are consistent. Two policies conflict if one policy allows access to certain data while the other one prevents access. The relationships among the various data must also be considered in this regard. For example, if access to generalized data is prohibited to some given entity, but access to more specific data is allowed, it represents a conflict. A formal analysis will help expose such inconsistencies using which errors can be corrected or appropriate conflict resolution policy will be chosen. For consistency purposes, it is necessary to ensure that the formal policy model complies with the underlying rules and regulations, expressed in natural languages. In this regard we also need to ensure that the enforcement rules comply with the formal models [12]. This can be challenging when the policies are enforced through algorithms implemented on devices having different form factors and capabilities, and we need to formally prove that these enforcement rules conform to our policy models. More research is needed in this area.

Note however, that policy analysis is not just an one-time effort. Policies are subject to change, attributes of users and data may change, and the context of access may change. Our framework should be able to support such evolution by understanding the nature of such changes and their impacts on the system. The support for dynamic policies [27], where policies are changed while they are deployed is essential in such cases. For example, if a caregiver is found to be untrustworthy, then his access must be revoked immediately even though he may be performing some tasks. Most models check for access before execution of an operation; if access privileges are changed while operations are executing no action is taken. A policy framework that supports dynamic policies where the type of change can be automatically analyzed to decide as to whether to continue or abort the operation is needed [28].

Evaluating risk of data access: In a dynamic and context dependent environment, it may be impossible to foresee all the situations in advance. Thus, a risk estimation strategy is needed for a risk-based access control approach. Access decisions need to be based on the risk associated with either granting or not granting the request. This risk is computed by analyzing the situation at hand, and taking into consideration contextual parameters such as the requestor's credentials, current access rights, history of past actions, etc. Together with the domain experts, we are working on identifying the relevant informative features. Most prior work on risk-based access control attempts to quantify the risk when the decision is known in advance [3, 8, 22, 29]. In this case, the correct decision may be unknown. Hence our approach is to investigate how the access control decision can be extrapolated in the presence of incomplete information and give a measure of risk if we make a wrong decision. A key line of work is focused on quantifying the associated risk of providing access through ML based classification models. Here, it is critical to extract the most appropriate features from existing data, which may require different transformations, such as counting the number of successful access requests in the past, categorizing user credentials, etc. There are also many different techniques for classification which have differing tradeoffs in terms of accuracy, scalability, and utility. The prediction probability obtained from each classifier can be considered as evidence and an overall risk score by using Dempster-Shafer's theory of belief. To avoid issues of algorithmic accountability, the contextual access control approach needs to provide justifications of why access is being granted or denied (e.g., change of user credentials, change in contextual parameters, etc.). In this context, this evaluation process is inter-related to the research in bias and fairness of AI/ML algorithms discussed in Section 4.3.

Analyzing access and data usage for potential abuse and attack: It is also important to evaluate the potential for abuse of released data (see, for example, report of elderly people's data being advertised as "These people are gullible. They want to believe that their luck can change," [9]). This requires a privacy-preserving anomalous activity detection system. At the lower level, based on history, only certain activities that could be potentially anomalous will be reported to the anomaly detection system. For example, for the vulnerable elderly population, any payment to non-white-listed accounts (e.g., utility companies could be whitelisted) and/or above a certain threshold (e.g., any payment above $1000) could be reported. This way only a limited amount of data would be disclosed for anomaly detection purposes. Once the potentially anomalous activity is reported, a ML model can further analyze the activity. Since the data available for building models for the vulnerable population is limited, we need to explore the setting where a generic anomaly detection built for the general population could be tailored for the vulnerable population. More specifically, we need to explore how recent advances in transfer learning could be applied in this domain, thus limiting the amount of vulnerable population data needed to train the model.

Adaptable policy configuration: To provide secure and privacy-preserving services to vulnerable populations, it is important that only legitimate access is granted at the right level of granularity. While this sounds like a traditional access control problem, there are many additional challenges in the context of vulnerable populations that require solutions beyond traditional access control. The key problem is that a static predetermined access control policy may not be easily formulated or even appropriate in such situations. Furthermore, it is difficult, if not impossible, to explicitly lay out all the possible access control decisions at the start, especially at the different data granularities. Here, it is necessary to develop an easy, automatically configurable access control system that can identify and ensure context dependent security constraints to ensure safety and security. It should be possible to build on prior work for dynamic coalitions [37]. Formal access control models operate under the principle that a user's request to a specific resource is honored if there is an explicit policy specifying that the user can access that resource. However, it is not feasible to explicitly specify a complete security policy. In such cases, access should be allowed prospectively based on the "need-to-know" requirement of a specific situation and based on the history of past actions. Alternatively, while legitimate access requests are allowed, they are retroactively examined to assess the legitimacy of the action through accountability/auditing measures. Here, since human resources are limited, inappropriate accesses need to be prioritized to enable quick investigation. Both approaches enable different security/utility/efficiency tradeoffs which can be appropriate for different contexts.

4.2 Utility-Driven Adaptive Policies

Ensuring privacy in the context of vulnerable population poses unique challenges – too little privacy protection can potentially dis- courage the vulnerable groups to communicate about their situations due to their privacy and safety concerns, while too much privacy (e.g., large noise added due to a privacy mechanism) would make vulnerable populations invisible to the world and to those who would help – creating a lack of visibility.

Privacy solutions, today, lie at an extreme on the privacy axis at the cost of final benefits offered to the stakeholder. In contrast, our goal is to provide an adaptive, context driven framework that emphasizes overall benefit to the vulnerable population and finds a reasonable trade-off for privacy. This idea is illustrated in Figure 3.

To that end, we need solutions that take the context of vulnerable populations into account. For example, referring back to the example of elderly lady Alice living alone, the classifier in Alice's (elderly person) smart assistant determines that Alice has fallen based on a limited set of sensory data, to preserve her privacy, but with low probability. So, the smart assistant probes more intimate sensory data to be sure that Alice has, indeed, fallen to determine if the medics should be called. However, if with the limited data, the classifier could confidently determine that Alice had not fallen, the probe of additional sensory data is not required. The realization of such an adaptive approach raises many challenges on how privacy should be defined as unlike differential privacy (DP) characterized by a single privacy parameter, in this context, different data may have associated different privacy parameters. This multiplicity of parameters is based on the level of privacy a) one could afford while meeting utility guarantees based on adaptive queries, b) be a result of the sensory data quality that varies from one client to the other, c) privacy preferences elicited by each user and so forth. Prior work on one-sided differential privacy [16] for example, provides yet another new model that exploits partial sensitivity of data to support increased utility (by allowing more data to be shared) while ensuring strict privacy properties on pre-identified sensitive parts of the data. The adaptive nature of our approaches thereby allows for integration of privacy/security into a framework that has an observe-analyze-adapt loop. Towards our goal of achieving context-aware adaptive privacy, we need to explore numerous complementary approaches identified below starting with producing new privacy models relevant for vulnerable populations. We discuss how to use these new privacy models to develop privacy-risk aware ML framework. Finally, we discuss how these ML models are to be leveraged to develop minimally invasive monitoring for vulnerable populations.

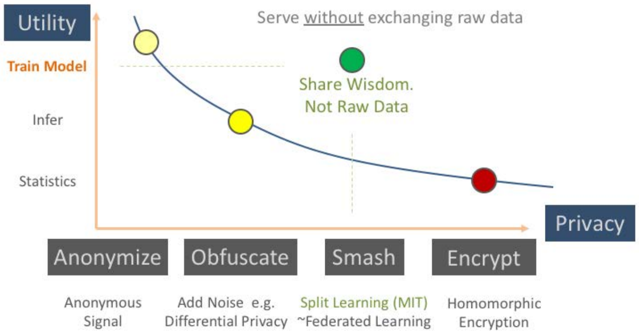
Utility
Train Model
Infer
Statistics
Serve without exchanging raw data
Share Wisdom. Not Raw Data
Privacy
Anonymize
Obfuscate
Smash
Encrypt
Anonymous Signal
Add Noise e.g. Differential Privacy
Split Learning (MIT) ~Federated Learning
Homomorphic Encryption

Figure 3: The data privacy versus data utility tradeoff

Developing enhanced model for privacy of vulnerable populations: The risk to privacy is higher for the vulnerable population as it is exceedingly difficult for them to understand the impact of their privacy-related options/decisions on their data privacy and security; and hence, become more susceptible to privacy violations. Differential privacy (DP), which is the state-of-the-art model for privacy, is not always a good fit for settings where the accuracy of sanitized data is exceedingly important for the healthy survival of vulnerable population. This is becuase DP involves introducing noise to the base data to sanitize it, which is why in the accuracy of such applications suffers immensely.

Let us say we have an algorithm $A : D \longrightarrow \{0, 1\}$ to detect if an event E (e.g., seizure) has occurred (i.e., 1) or not (i.e., 0). The algorithm makes this decision based on the given data $x \in D$ (i.e., the set of all databases). The data contained in x is sensitive, and we want A to be privacy-preserving (i.e., it protects privacy of the sensitive information) without compromising on the accuracy of the detection.

Differential privacy (DP) is the state-of-the-art model for privacy but it is not a good fit for such settings because under DP the accuracy of such applications suffers immensely. Thus some relaxations of DP have been proposed to fine-tune the the utility-privacy trade-off in the favour of detecting E. To fine-tune such trade-offs, one-sided differential privacy [16] exploits the partial sensitivity of the data, protected differential privacy [15] uses the label of a record, and sensitive privacy [2] employs outlyingness of records. When a record's sensitivity can be defined based on itself (i.e., it can be defined without the other records), then one-sided DP and protected DP can admit privacy-preserving mechanisms that have practically meaningful accuracy. However, when the sensitivity of a record is determined by the database it belongs to, it is necessary to resort to sensitive privacy, or generalize some of the notions (e.g., neighboring databases) in one-sided and protected DP.

In some situations, however, it is possible that a combination of the two (or more) different mechanisms—which are private under different notions—provides the desired utility-privacy trade-off. For instance, computing a final output using two partial results, one obtained via a sensitively private mechanism and the other using a one-sided DP mechanism, gives better accuracy. Thus, we will build a privacy model that enables us to analyze the problems for this setting and is able to quantify privacy under such a hybrid model of privacy. For this we plan to use tailored differential privacy [18], which provides a way to quantify and analyze the privacy of such hybrid models.

Privacy Risk-aware ML framework for vulnerable population: Over the years, many privacy attacks against machine learning (ML) models ranging from model inversion to membership inference have been launched [13]. Using differential privacy emerged as the main protection against these attacks. Although differential privacy mechanism provides important protections, empirical evaluations have shown that, in many cases, the ϵ value required for protecting against a certain attack (e.g., model inversion attack where sensitive attribute is predicted) may totally kill the utility of the ML model [10]. Therefore, finding the privacy parameter (e.g., ϵ in differential privacy or other privacy-parameters for a new privacy model like the one suggested in the previous task) that provides desired utility and adequate protection against attacks may not be always feasible.

We envision a complementary approach that is based on feasible attack based risk modeling against vulnerable population and incorporating pre-processing (e.g., modify data before it is used by differentially private or any other privacy definition supporting learning mechanism) and post-processing techniques (e.g, provide a wrapper for a given classifier to reduce model inversion attack) combined with carefully selected privacy parameters (e.g. ϵ for differential privacy) to get good utility and adequate protection against realistic privacy attacks. Such attack modeling based risk assessment are in common use in several security domains.

Minimally invasive monitoring: Decision support technology often uses machine learning models (such as classification) to analyze collected data to support both short term interventions, such as, in Example 1, Alice's smart assistant may have a classifier to detect falls and call the paramedics if required, to long term adaptations, for example, Alice's smart assistant can find trends in Alice's general health and recommend her when it might be time to go to the doctor. Even though more accurate ML models could potentially be built using our privacy-risk aware ML framework above, still they do not provide any guarantees on the quality of data output. Decision support tasks require guarantees on the quality of the output, especially for false negatives that may prevent timely interventions. Decision support systems for the vulnerable population pose an interesting dichotomy for privacy preserving technology: Decision confidence and privacy put contrasting requirements on the amount of data released.

Our work explores a radically new concept of *minimally invasive monitoring* (MIM) that attempts to resolve the above paradox. The envisioned MIM approach changes the objective to achieve a (probabilistic) bound on utility while optimizing (maximizing) privacy, instead of just optimizing for utility all the while implementing strong privacy guarantees (as is done traditionally). The utility constraint, itself, is set in a conservative manner such that the decision support task results in a limited level of false negatives. Decision support tasks such as classification or queries can often be implemented in a manner such that false negatives can be arbitrarily decreased at the cost of increasing false positives. Such a strategy of using generalization/specialization to control the tradeoff between false negatives and positives is broadly applicable in classifiers. The envisioned approach exploits such an observation to support guaranteed utility while maximizing privacy. A MIM framework can, thus, be viewed as a progressively invasive system that explores data in the context of a monitoring task through a coarse filter with an elevated level of privacy but explores the data using a finer filter more invasively only if it passes through the coarse filter. Consider, again, the scenario in Example 1. The classifier in Alice's smart assistant determines that Alice has fallen based on a limited set of sensory data, to preserve her privacy, but with low probability. So, the smart assistant probes more intimate sensory data to be sure that Alice has, indeed, fallen to determine if the medics should be called. However, if with the limited data, the classifier could confidently determine that Alice had not fallen, the probe of additional sensory data is not required. The realization of the MIM raises a large number of challenges -– how should privacy be defined since unlike DP characterized by a single privacy parameter, in MIM, different data may have associated different privacy parameters based on the level of privacy we could afford while meeting utility guarantees; how do we support MIM algorithmically so as to get the requisite utility while ensuring privacy; how do we ensure fairness and absence of bias since in MIM different individuals may be monitored at different levels of privacy based on the needs of the task. Furthermore, since MIM allows adaptive invasive exploration, it could be susceptible to misuse. Mechanisms to implement accountability via evidence demonstrating need for invasive exploration must be supported to build in checks and balance in the system.

4.3 Fair and Unbiased Big Data, AI and ML

SAFE-PASS employs cutting edge AI/ML techniques to deliver assistive services to vulnerable populations. Thus one of our goals is to address problems of implicit biases in data collection, extraction, fusion, model learning, and analytics that could lead to malicious discrimination, unfairness, misrepresentation, denial of access to services and/or inadequate engagement. In addressing fairness and bias, the first key question is to identify attributes that are sensitive and need to be protected. In our context, such attributes could include features such as race, social status, economic status, level of active involvement of family in a person's well-being, etc. A fair algorithmic solution must ensure that such factors do not influence the machine learning models learnt, or decisions taken, in a discriminatory manner. Further, there is a need to design fairness mechanisms that provide equitable performance without the need to explicitly identify certain people as trafficking victims.

While designing fair and ethical approaches in algorithmic decision making have gained significant ongoing research attention [20], supporting the PASS technologies in the context of the vulnerable population raises several new challenges. For instance, a key to building trustworthy machine learning solutions is explainability and transparency. We need to also consider explainability in a broader sense of explaining decisions to potentially cognitively impaired due to their vulnerability. The concept of explainability also spreads to system and model developers to ensure that the models being learnt do not inadvertently include biases such as bias based on cognitive level of the individual. Another critical challenge is to design monitoring systems to detect bias and/or ensure fairness when models are applied to dynamically incoming data. This is especially pertinent when dealing with vulnerable populations as the operating conditions may change rapidly due to changes in health status, cognitive ability, and threats from external agents. Hence, the provenance of the models/decisions needs to be maintained and the data about the environment and individuals embedded in the environment to be captured. For instance, to ensure that the assistive technologies/models are not being misused and the models/ algorithms do not veer away from being fair, we need to capture data about the environment to verify desirable properties against the models. This raises a new challenge of "monitoring" the (fall) "monitoring system" that, unless done carefully, could catapult into potential issues of privacy and misuse. Another challenge is model adaptation that is inherent in dynamic systems. Such adaptations must not re-introduce bias. Below we discuss some key challenges/directions that needs to be explored.

Defining fairness in the SAFE-PASS setting: Specifying what is fair in our context is non-trivial and challenging: existing definitions of fairness are often driven by application-specific and even legal considerations. There are numerous fairness definitions and bias mitigation algorithms proposed in the literature [20], and new ones continue to emerge. These notions can be roughly categorized into demographic aware, error aware, impact aware. Error aware definitions can be applied when the focus is on achieving similar error rates for diverse groups, and that these errors should be minimized. Demographic aware definitions can be applied in situations where all demographic groups need to be represented equally or proportionally in an outcome or decision. Impact aware definitions are those that incorporate the long-term impact of a decision. A data-driven policy may lead to a different long-term impact for different sub-populations. However, common notion of fairness is not expressive enough to capture constraints [14, 36] and requirements specific to the contexts in which ML algorithms are deployed. To address the challenges, we need to develop methods and tools tailored to vulnerable populations to elicit information about the conception and reception of utility and fairness.

Challenges due to data biases: Data-driven solutions to assist vulnerable groups rely on data integrated and fused from all types of data, such as electronic health records, multiomics data, public health data, insurance claims data, social media data, and non-traditional data collected from wearable devices, smartphones, GPS, and satellite data. These data sources are prone to different forms of data biases, such as cognitive bias, measurement error and misclassification, historical bias, and missing data, that disproportionally affect vulnerable and minority groups. Furthermore, data collected from these sources may suffer from selection bias and imbalance due to differences in subpopulations in terms of access to resources or geographic limitations, and thereby may not be representative of vulnerable groups.

Even if the data is unbiased, there is no guarantee that the algorithm or downstream application will be unbiased. Biases can be introduced by handling and transforming data throughout the ML pipeline and during model development and deployment. Our research explores methods that holistically address several types of data biases pertinent to vulnerable groups. Specifically, we need to develop automated diagnosis tools for data biases that appeal to how the data were collected, integrated, and processed through the ML pipeline. In this direction, we need to develop efficient algorithms to detect and automatically correct several types of data biases such as selection bias, data imbalance, measurement bias and missing data. We are exploring provenance techniques to monitor and inspect training data for various forms of bias issues that might be introduced in various stages of ML pipeline, and to trace back fairness and bias issues of downstream ML models to training data and decisions made during the ML pipeline that led to the bias.

Explainability leads to better adoption: Transparency, understandability and explainability are integral in the design and widespread adoption of data-driven tools to assist vulnerable groups. They help to build trust among different stakeholders and are often required by law also. In our setting, this is more complex since we are dealing with different type of stakeholders - doctor, nurse, caregiver, family members, or vulnerable individuals themselves that could be elderly or mentally impaired - that may have different perception of trust and risk. In such situations, building trust through explainability can be significantly complex since trust depends on understanding and respecting the needs and interests of trustees. Therefore, we require methods that provide the types of explanations that are most relevant to different stakeholders in specific societal settings. There has been a recent resurgence of interest in explainable artificial intelligence (XAI) [11, 21] that aims to reduce the opaqueness of AI-based decision-making, which aims to provide human-understandable explanations of outcomes or processes of algorithmic decision-making systems. Most of the existing methods in XAI focus on the attribution of responsibility of an algorithm's decisions to its inputs, by ranking input features based on their importance for a particular decision made by an algorithm. However, these methods can produce incorrect and misleading explanations primarily because they focus on the correlation between the input and output of algorithms as opposed to their causal relationship. We need to develop novel methods for generating efficient and reliable causal and contrastive explanations for ML models developed for assisting vulnerable population. These models should not just predict which individual is vulnerable, but also provide recommendations for how to change one's risk trajectory toward a better outcome. Generating such explanations in our context is particularly challenging, and it requires careful examination of feasibility and effectiveness of the generated explanations, as well as privacy and security considerations.

Robustly maintaining fairness in the presence of concept/ distribution drift: ML algorithms in practical settings require adaptation based on distribution shifts over time. This is especially pertinent in the context of vulnerable populations where data about physical characteristics (e.g., loss of hair during chemotherapy), cognitive abilities, or housing location (e.g., for safety in trafficking scenarios) may change rapidly. Models learnt, while fair on the original data, may need to be adapted to continue to provide desired properties. We plan to explore algorithmic innovations to dynamically adapt SAFE-PASS algorithms to ensure that the system does not veer off too far from its desired state.

4.4 Adaptable and Scalable Protocols and Technology

In developing a technology infrastructure for SAFE-PASS, we observe that support for adaptivity is a primary design criterion. The challenge at hand is to design an adaptive architecture that supports these composite needs of privacy, utility, accountability, fairness and explainability. Towards this end, we explore the "Observe - Analyze - Adapt" (O-A-A) loop in computational reflection [19, 34, 35] as a principled approach for designing an adaptive middleware and data management architecture to capture the dynamic security and privacy needs of SAFE-PASS applications.

We envision the realization of the SAFE-PASS framework as a middleware or metasystem (see Figure 4) that intercepts information flow from heterogeneous data sources at the base-layer to applications and users/proxies at the consumer-layer to implement the privacy/security needs of the observed entities, as well as to meet the required legislative mandates. In the envisioned use cases for SAFE-PASS technology, data obtained from heterogeneous information sources via the Observe/Monitor Module) are integrated and undergoes several layers of transformation to create semantically meaningful observations to drive applications. As information is exchanged from data sources to consumers, SAFE-PASS mechanisms will need to be automatically triggered enroute to support fairness/privacy/security guarantees. In SAFE-PASS, privacy techniques (e.g., OSDP, MiM, encryption) are implemented across multiple components of the O-A-A architecture to perform (logical) adaptation of what information is collected and how it is processed/shared. In short, SAFE-PASS middleware will implement technology modules that realize "PASS" to ensure "SAFE" information flows. SAFE-PASS takes a novel approach in how the reflective architecture is exploited. In addition to enabling logical adaptation of data collection and actuation at the base level based on policies, the SAFE-PASS Adaptation Engine enables adaptation at higher meta-layers by adapting and appropriately choosing the techniques that govern information flow. This higher-level adaptation includes the choice of privacy policies themselves, parameters associated with the privacy mechanisms, trust boundaries, risk/utility margins and bias tolerance limits. The goal is to provide a holistic adaptation that can support human overrides, nudges, and recommendations to balance risk tolerance vs. application utility in a context-sensitive way.

The architecture provides mechanisms for reactively triggering new SAFE-PASS methods on the fly. The adaptation engine safely steers the privacy strategies to address the utility/risk tradeoffs specified by users. The middleware allows for adaptations to be explored/initiated at the logical level; this enables on-the-fly validation of the feasibility/value of the changes induced by the mechanisms prior to enforcing them into the physical systems. Additional modules address the SAFE-PASS goals of fairness, explainability and accountability. The analyze/process components encapsulate bias detection and mitigation techniques, methods for ensuring fairness in data-driven decision-making algorithms, and methods for generating explanations for end-users and system developers.

With dynamic adaptation, accountability and auditability are essential to establish provenance. This is particularly true when sensitive information revealed at times of need (e.g., health emergency, natural disaster) can be misused later. Audit logs recording accesses to data and adaptations performed will be stored and shared with trusted audit authorities. SAFE-PASS will adopt blockchains and smart contracts technology to record sharing / usage data. One of the major concerns when using blockchains is that of latency and scalability [17, 39]. This is a major research challenge in the context of SAFE-PASS. We are exploring a hybrid approach to combine blockchain and traditional databases.

5 DISCUSSIONS AND CONCLUSION

The past two decades have witnessed unparalleled advances in technology: People connected to each other, network in the hands of everyone, and knowledge and services on demand. More recent advances in sensing and actuation and ambient intelligence via big data analytics (BD), artificial intelligence (AI) and machine learning (ML) hold enormous potential to benefit vulnerable groups. At the same time, the untrammelled and unrestricted use of the same technology, coupled with issues related to flaws, and algorithimic biases and fairness, can equally cause enormous harm to the same groups. In this paper, we present the SAFE-PASS vision: A societally responsible, secure and trustworthy cyberspace that puts algorithmic and technological checks and balances on the indiscriminate sharing and mining of data. SAFE-PASS provides a unique opportunity to establish trust in technology.

The core design principles of SAFE-PASS is a novel data integration, analysis and sharing paradigm which we call Selective Secrecy and Structural Transparency. Selective Secrecy results in judiciously providing strong levels of security and privacy to shared data by default and updating the levels based on situational awareness and an evaluation of the utility of the sharing. Structural Transparency results in enabling questions such as, "what information about me is out there," "am I being misidentified or miscategorized," "is my information being used against me," or "is the data requested

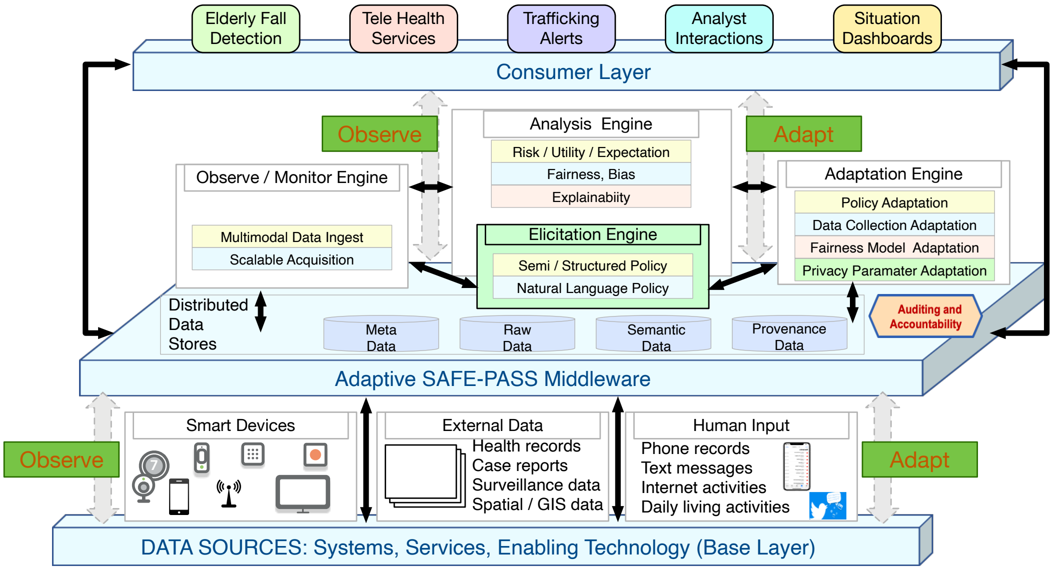

Figure 4: SAFE-PASS technology realization architecture

too invasive," to be asked and answered and preventive actions to be taken.

In this paper, we discussed some of the core security and privacy challenges that we are trying to address in bringing the SAFE-PASS vision to fruition through our research. Our research touches upon five major thrust areas: (i) Expanding knowledge and understanding of security and privacy perceptions and expectations in vulnerable groups, which significantly contribute to their unwillingness to share data, and use that knowledge to drive research in (a) mitigating missing/imbalanced data problems, (b) understanding and modeling security and privacy risks of data sharing, and (c) modeling utility of data sharing. (ii) Developing a risk-adaptive, policy model capable of capturing and articulating security and privacy expectations of users that are relevant in a particular context and develops associated technology to ensure provenance and accountability. (iii) Developing robust AI/ML algorithms that are transparent and explainable with respect to fairness and bias so that these techniques reduce/eliminate discrimination, misuse, privacy violations, or other cyber-crimes. (iv) Developing models and techniques for a nuanced, contextually adaptive, and graded privacy paradigm that allows trade-offs between privacy and utility and (v) Developing adaptable and scalable systems and technology to support the SAFE-PASS paradigm.

We acknowledge that new algorithms and technology that are needed to bring SAFE-PASS to fruition can potentially be exploited by adversaries to cause harm to the same vulnerable groups that SAFE-PASS is intended to protect. For example, selective secrecy, which is a core tenet of SAFE-PASS, implies relaxing confidentiality or privacy when the utility of data sharing is higher. Thus, an attacker can manipulate a scenario where the perceived benefits are higher than they are and cause sensitive informaition to be revealed. Or, for example, explainable AI/ML techniques for SAFE-PASS can provide more food for thought to attackers than they would otherwise have. Nonetheless, the benefits that SAFE-PASS can potentially bring in to society by strengthening the communication channel between science and technology and the vulnerable groups so that this segment of the society better understands how technology both benefits them and protects them without the fear that the technology would harm them, far outweighs the risks posed by the SAFE-PASS paradigm. SAFE-PASS presents a unique opportunity to re-invigorate the social contract that our forefathers wrote for a representational democracy.

ACKNOWLEDGMENT

We gratefully acknowledge the funding from the following agencies and organizations: the U.S. National Science Foundation (grant #1822118, #2226232, #OAC-1828467, #OAC-2115094), the member partners of the NSF IUCRC Center for Cyber Security Analytics and Automation – AMI, NewPush, Statnett, Cyber Risk Research, NIST and ARL – the State of Colorado (grant #SB 18-086) and the authors' institutions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, or other organizations and agencies.

REFERENCES

[1] Manar Alohaly, Hassan Takabi, and Eduardo Blanco. 2019. Automated Extraction of Attributes from Natural Language Attribute-Based Access Control (ABAC) Policies. *Cybersecurity* 2, 1 (2019), 1–25.

[2] Hafiz Asif, Periklis A Papakonstantinou, and Jaideep Vaidya. 2019. How to Accurately and Privately Identify Anomalies. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 719–736.

[3] John A Clark, Juan E Tapiador, John McDermid, Pau-Chen Cheng, Dakshi Agrawal, Natalie Ivanic, and Dave Slogget. 2010. Risk Based Access Control with Uncertain and Time-Dependent Sensitivity. In *International Conference on Security and Cryptography*. IEEE, 1–9.

[4] Council for International Organizations of Medical Sciences. 2016. International Ethical Guidelines for Health-Related Research Involving Humans. Available at https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf.

[5] Department of Homeland Security. [n. d.]. Human Trafficking Laws & Regulations | Homeland Security. https://www.dhs.gov/human-trafficking-laws-regulations

[6] Department of Justice. [n. d.]. Elder Abuse and Elder Financial Exploitation Statutes | EJI | Department of Justice. https://www.justice.gov/elderjustice/prosecutors/statutes

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Nathan Dimmock, Jean Bacon, David Ingram, and Ken Moody. 2005. Risk models for trust-based access control (TBAC). In *International Conference on Trust Management*. Springer, 364–371.

[9] Charles Duhigg. 2007. Bilking the Elderly, With a Corporate Assist. *New York Times* (May 20 2007). https://www.nytimes.com/2007/05/20/business/20tele.html

[10] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. ACM, NY, USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[12] Andy Chunliang Hsu and Indrakshi Ray. 2016. Specification and Enforcement of Location-Aware Attribute-Based Access Control for Online Social Networks. In *International Workshop on Attribute Based Access Control*. ACM, 25–34.

[13] Bargav Jayaraman and David Evans. 2019. Evaluating Differentially Private Machine Learning in Practice. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 1895–1912. https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman

[14] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2021. An Algorithmic Framework for Fairness Elicitation. In *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference (LIPIcs, Vol. 192)*, Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2:1–2:19. https://doi.org/10.4230/LIPIcs.FORC.2021.2

[15] Michael Kearns, Aaron Roth, Zhiwei Steven Wu, and Grigory Yaroslavtsev. 2016. Private Algorithms for the Protected in Social Network Search. *Proceedings of the National Academy of Sciences* 113, 4 (2016), 913–918.

[16] Ios Kotsogiannis, Stelios Doudalis, Samuel Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. 2020. One-Sided Differential Privacy. In *36th IEEE International Conference on Data Engineering (ICDE)*. IEEE, 493–504.

[17] Tsung-Ting Kuo, Hyeon-Eui Kim, and Lucila Ohno-Machado. 2017. Blockchain Distributed Ledger Technologies for Biomedical and Health Care Applications. *Journal of the American Medical Informatics Association* 24, 6 (2017), 1211–1220.

[18] Edward Lui and Rafael Pass. 2015. Outlier Privacy. In *Theory of Cryptography Conference (TCC)*. 277–305.

[19] Pattie Maes. 1987. Concepts and Experiments in Computational Reflection. *ACM Sigplan Notices* 22, 12 (1987), 147–155.

[20] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[21] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[22] Ian Molloy, Luke Dickens, Charles Morisset, Pau-Chen Cheng, Jorge Lobo, and Alessandra Russo. 2012. Risk-Based Security Decisions under Uncertainty. In *Conference on Data and Application Security and Privacy*. ACM, 157–168.

[23] Masoud Narouei, Hassan Takabi, and Rodney Nielsen. 2018. Automatic Extraction of Access Control Policies from Natural Language Documents. *IEEE Transactions on Dependable and Secure Computing* 17, 3 (2018), 506–517.

[24] Mariam Nouh, Abdullah Almaatouq, Ahmad Alabdulkareem, Vivek K Singh, Erez Shmueli, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. 2014. Social Information Leakage: Effects of Awareness and Peer Pressure on User Behavior. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 352–360.

[25] Jinkyung Park, Eiman Ahmad, Hafiz Asif, Jaideep Vaidya, and Vivek K Singh. 2022. Privacy Attitudes and COVID Symptom Tracking Apps: Understanding Active Boundary Management by Users. In *17th International Conference on Information iConference 2022: Information for a Better World: Shaping the Global Future (Lecture Notes in Comouter Science, Vol. 13193)*. Springer.

[26] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics* 34, 2 (2008), 257–287.

[27] Indrakshi Ray. 2004. Real-Time Update of Access Control Policies. *Data Knowl. Eng.* 49, 3 (2004), 287–309.

[28] Indrakshi Ray. 2005. Applying Semantic Knowledge to Real-Time Update of Access Control Policies. *IEEE Trans. Knowl. Data Eng.* 17, 6 (2005), 844–858.

[29] Farzad Salim, Jason Reid, Ed Dawson, and Uwe Dulleck. 2011. An Approach to Access Control under Uncertainty. In *International Conference on Availability, Reliability and Security*. IEEE, 1–8.

[30] Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. "They Basically like Destroyed the School One Day" On Newer App Features and Cyberbullying in Schools. In *20th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. 1210–1216.

[31] The European Union. 2018. What is GDPR, the EU's New Data Protection Law? https://gdpr.eu/what-is-gdpr/.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in neural information processing systems*. 5998–6008.

[33] Nalini Venkatasubramanian, Mayur Deshpande, Shivajit Mohapatra, Sebastian Gutierrez-Nolasco, and Jehan Wickramasuriya. 2001. Design and Implementation of a Composable Reflective Middleware Framework. In *Proceedings 21st International Conference on Distributed Computing Systems*. IEEE, 644–653.

[34] Nalini Venkatasubramanian, Carolyn Talcott, and Gul A Agha. 2004. A Formal Model for Reasoning about Adaptive Qos-Enabled Middleware. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 13, 1 (2004), 86–147.

[35] Nalini Venkatasubramanian and Carolyn L Talcott. 2001. A Semantic Framework for Modeling and Reasoning about Reflective Middleware. *IEEE Distributed Systems Online* 2, 06 (2001), null–null.

[36] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why Fairness Cannot be Automated: Bridging the Gap between EU Non-Discrimination Law and AI. *Computer Law & Security Review* 41 (2021), 105567.

[37] Janice Warner, Vijayalakshmi Atluri, Ravi Mukkamala, and Jaideep Vaidya. 2007. Using Semantics for Automatic Enforcement of Access Control Policies among Dynamic Coalitions. In *12th ACM Symposium on Access Control Models and Technologies (SACMAT)*. 235–244.

[38] Xusheng Xiao, Amit Paradkar, Suresh Thummalapenta, and Tao Xie. 2012. Automated Extraction of Security Policies from Natural-Language Software Documents. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*. 1–11.

[39] Hongru Yu, Haiyang Sun, Danyi Wu, and Tsung-Ting Kuo. 2019. Comparison of Smart Contract Blockchains for Healthcare Applications. In *AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association, 1266.