Research Statement

Sarah Scheffler

Cryptography has grown from a specialized niche tool into a ubiquitous technology—and as a result, policymakers now often grapple directly with questions of security, privacy, and transparency. My research provides two complementary sets of tools to these policymakers: First, through a mixture of policy analysis, technical frameworks, and empirical legal research, I directly investigate the difficult tradeoffs between the goals of harm reduction, auditability, and privacy in encryption policy, especially for end-to-end encrypted social media settings. Second, with this analysis guiding my design, I build applied cryptography protocols that improve upon the state-of-the-art in accountability, privacy, and efficiency, widening the set of options for addressing the policy problem. Throughout both sides of my research, I engage directly with policymakers in industry, government, and academia, using my cryptography expertise to build a better future.

Content moderation for end-to-end encryption

A growing number of communication services are offering *end-to-end encryption* (E2EE) in which only the users have the keys, and the service providers cannot read or tamper with the plaintext content. This is a win for privacy and security for average people—but brings with it governance challenges, as encryption makes it harder to catch misbehaving users of these E2EE services. In particular, E2EE disrupts many naive content moderation systems that the provider might have performed to detect spam, malware, child safety violations, hate speech, or other problematic content. On the other hand, naive "privacy-preserving" content moderation systems capable of functioning within E2EE could pose significant risks to freedom of expression, privacy, and security. My research tackles this problem from all sides, with a special focus on transparency mechanisms that roadblock the slope to censorship.

Improving transparency and auditability. The year 2021 was a flashpoint in the global policy debate over content moderation in end-to-end encryption, with Apple's proposal of a hash matching system which would automatically alert Apple if an iCloud Photos user uploaded content that shared a hash with known Child Sexual Abuse Material (CSAM), while preserving the privacy of non-matches. Apple delayed the system indefinitely in the face of criticism of the system's risks to free speech, privacy, and security, but governments took note of the possibility of proactively detecting CSAM, terrorist, or other content under encryption.

In recent work to appear at the upcoming IEEE Security & Privacy 2023 (a tier-1 computer security venue) [SKM23], I identified and implemented technical improvements to the transparency and auditability of such a system. To *build trust in the implementation* we created a system that enforces notification to users if their content was revealed to Apple, after a delay allowing the moderator to process the detection in some way (e.g. passing it to law enforcement). We implemented this system using the state-of-the-art malicious-secure authenticated garbling approach to multi-party computation of Wang et al. To *build trust in the hashset* we provide two contributions. The first is a threshold signature scheme allowing child safety organizations to certify their part of the hash list in a publicly verifiable manner, while ensuring the list remains private to clients and robust against malicious attempts to alter the list. Second, we build a scheme that allows the central moderator to prove that specific hashes are *not* contained in the hashlist, providing credibility for claims that they are only using the moderation scheme for its intended purpose. To do so, we propose using a zero-knowledge proof of non-membership in a Cuckoo table of blinded hash values. Proving non-membership in a set could be accomplished by a negative accumulator, but this generic approach is inefficient, and does not make use of the existing public information in our specific content moderation setting. Our approach shrinks the computation time and communication required by making use of existing public information. By combining homomorphic commitments with the classic proof of knowledge of discrete

logarithm by Chaum et al., we create a proof of non-membership whose size is dependent only on the security parameter. These interventions combined raise both the technical and normative bars to misusing a content moderation system for more censorious purposes.

Understanding the landscape. In [SM23], I completed a Systematization of Knowledge (SoK) on content moderation for end-to-end encryption which was accepted at the Privacy Enhancing Technologies Symposium 2023, a top venue for the specific subject matter of privacy. I also presented an early version of this work at the DIMACS Computer Science and Law workshop on Content Moderation in May, 2022.

To do the first part of this work, I performed a massive literature analysis touching thousands of papers and diving deep into about 120 industry and academic proposals for content moderation under E2EE. My analysis unifies the privacy-preserving content moderation literature and includes not only the oft-discussed policy challenges of child safety and disinformation, but also topics that were nearly absent from the current debate, like corporate and parental monitoring of encrypted internet traffic.

In the second part, I provide much-needed contextualization of this existing technical research on content moderation in E2EE. I provide a general framework to analyze design choices in privacy-preserving content moderation that is useful for policymakers, service providers, technologists, and civil rights advocates alike. I go into detail on the current set of options for content moderation under encryption, including the level of privacy, detection mechanism, cryptographic tools used, security guarantees, efficiency, and any transparency properties the system offers. This work crystallizes the importance of a deep technical and policy understanding of content moderation in E2EE – there are already deployments of E2EE content moderation, and they do in fact suffer from misuse and lack of transparency.

Engaging with policymakers. As a part of this research agenda, I briefed a group of engineers and decision-makers at Apple, including privacy chief Erik Neuenschwander, to provide feedback on the design of Apple's proposed content moderation systems. Separately, I have actively participated in formal discussions of the U.K.'s Safety Tech Challenge Fund, an initiative by the U.K. Home Office to build end-to-end encrypted systems that moderate content for child safety goals. I submitted a formal written comment [SMK22] during the Challenge Fund's evaluation process which argued for improved transparency, abuse resistance, and accuracy.

Ongoing work. I have already begun two additional lines of work on this topic: First, I turn to one of the most pressing motivations for moderating content under encryption: to preserve the safety of children against various forms of sexual abuse. Although this topic is at the forefront of the debate over legislating access to plaintext in democratic governments, surprisingly little research has been conducted on the scope of the problem, and the extent to which encryption does or does not stymie child safety efforts in the U.S. In ongoing empirical legal research, I am investigating public dockets to examine the role encryption plays in U.S. Federal District Court prosecutions of 18 U.S.C. 2252 and 2260, which prohibit possession and production of CSAM and some related harms against children. Second, I am researching the possibilities for cryptographic improvement of the transparency and integrity of the terrorism-related hash-sharing database of the Global Internet Forum to Counter Terrorism, in collaboration with a member of their Technical Working Group.

Cryptographic formalization of law and policy

Another line of my research seeks to deepen our understanding of the consequences of a particular law or policy by using cryptography.

My deepest work on the subject concerns an increasingly common question of U.S. law which years of litigation and pure legal scholarship have thus far failed to answer: *Can the government compel a device's owner to enter their password to decrypt their device?* I have two papers on the subject: The first [SV21][1] was published in Usenix Security 2021, (another top-tier conference in computer security) and formed the backbone of my Ph.D. dissertation. The second [CSV22][1] was published at the second ACM Symposium on Computer Science and Law (ACM CS&Law) in November, 2022. Both of these works resulted in formal technical models (based upon Interactive Turing Machines) of the legal doctrine governing compelled decryption. I then used those models to analyze several plausible compelled decryption scenarios. These

works revealed key differences in two different legal approaches to compelled decryption—both of which are consistent with prior non-encryption-related caselaw, but represent very different interpretations of what the Government may and may not compel. This divergence only became visible as a result of these technical models, thus illuminating an important gap that remains unaddressed by the pure legal literature.

In addition to helping answer legal and policy questions, these models can be used to investigate other questions of research. For example, among other findings, our formal analysis showed that the secret inputs to Secure Multi-Party Computation were particularly vulnerable to being compelled in a way that did not apply to other cryptosystems. In [SV21] I also considered the flip-side of the question: could I define and build a system without this weakness? I defined *FC-resilience* to capture this security property, and constructed an FC-resilient form of multi-party computation that I made available as open-source code [Scheffler21].

I have also completed other works that formalize law and policy. Most recently, in [STV22][1], I built a formal framework for analyzing whether one work is derivative of another for the purposes of copyright law, based around the computational complexity idea of description length; I also presented this work at ACM CS/Law 2022.

Ongoing work. I plan to extend my existing work on compelled decryption, both in research and in policy impact. As future technical research, our work in [CSV22][1] points to new opportunities for FC-resilient systems that I wish to unify with my existing work in this area [SV21, Scheffler21][1]. I also hope to extend our technical framework so that it can help answer a related difficult question of law: determining whether the testimony inherent in a given action action is *explicit* (analogous to answering a written or oral question) or *implicit* (i.e. what can be learned about a respondent's beliefs by their act of responding to a subpoena). I also have plans for active policy engagement in the future: First, building upon initial positive feedback from law professors, I am developing a law-first version of this paper to better communicate the technical ideas to a non-technical audience, so as to increase the impact and the potential for adoption by courts. Second, I plan to write an amicus brief when this topic inevitably reaches the Supreme Court. My brief will illuminate the difference our framework discovered between the two approaches to the legal doctrine, and will argue for the consistency of one of the two models with the first principles of the Fifth Amendment.

Looking at other ongoing work on this area, I have begun collaborating with roboticists to update and formalize my 2019 work on autonomous weapon systems [SO19]. This work, which won second place in the inaugural ACM CS&Law's Student Paper Competition, pointed out discrepancies between theory and reality in the legal literature on autonomous weapons. In the future I plan to take this work significantly further to provide a firmer ground from which to analyze the legality of various forms of autonomous weapon systems under the Law of Armed Conflict, as part of a broader research thrust into private and transparent robotics.

Advances to the frontier of zero-knowledge proofs

None of the methods described in the previous section would be possible without making advances in applied cryptography itself. My second research direction is to make foundational advances in traditional applied cryptography. Zero knowledge proofs allow a "prover" to convince a "verifier" of the truth of an NP statement, without revealing the witness that shows why the statement is true; for example, one might prove that a circuit is satisfiable without revealing the satisfying input. They are also key components of multi-party computation, in which a collection of parties jointly compute the output to a function without revealing anything about their sensitive input data.

Both of these technologies are key components of the more socially beneficial uses of cryptography, Zero-knowledge proofs appear in use cases as far ranging as nuclear armament verification, crime scene DNA non-matches, and enforcing rules for data surveillance and warrants. Multi-party computation has a similarly impressive list of applications, including detecting tax fraud, privately computing prices in electricity markets, avoiding satellite collisions, and measuring the gender and racial pay gap in the city of Boston.

My first improvement to zero-knowledge proofs was BooLigero [GSV21][1], an adaptation of the Ligero proof system for arithmetic circuits of Ames et al. to the Boolean setting which I presented at Financial

[1]The author order of these papers follows an alphabetical convention and order does not represent contribution.

Cryptography in 2021. For Boolean circuits like SHA-3, we achieved a reduction in proof size of $1.75 - 3\times$ over original Ligero without sacrificing prover or verifier runtime or memory. For hybrid arithmetic-Boolean operation circuits like SHA-2, we improved proof size by up to $1.6\times$.

I continued this line of research with a work at Applied Cryptography and Network Security 2021 [GHS+21][1]. Our new TurboIKOS system follows the "MPC-in-the-head" or "IKOS" framework of Ishai et al., in which the prover commits to emulated executions of several MPC parties, and then opens some of these to the verifier, who can use them to verify the prover's correct behavior with a probability based on the number of revealed parties. TurboIKOS remains state-of-the-art in MPC-in-the-head proofs. It improves upon Baum and Nof's prior work incorporating the Beaver triple "sacrificing" approach of MPC for use in zero-knowledge proofs, achieving a proof size comparable to the "cut-and-choose" approach of Katz et al. and the polynomial-interpolation approach of Baum et al., but with significantly lower memory costs.

I continue to research this line of work, investigating the use of puncturable pseudorandom functions to generate much of the info in the emulated MPC-in-the-head parties to further reduce the proof size.

Future work

Aside from privacy, the other tech policy elephant in the room is autonomous decision-making, especially systems based on artificial intelligence (AI) and machine learning. Policymakers around the globe are already grappling with setting regulatory requirements on current autonomous decision-making systems, but these challenges will expand tenfold in the upcoming age of popular robotics, from self-driving cars to drones to house-care robots. The existing technical frontier in this area shares many of the properties of the earlier traditional AI field: a lack of privacy, verifiability, and security. My future research agenda identifies gaps in these systems that cam be filled with cryptographic tools—especially in the privacy of multi-agent systems, and verifiability of robot behavior—and thus addresses policy challenges in many different domains, from civilian to commercial to military.

Privacy. There is a growing body of literature proposing approaches to measuring, preserving, and proving the privacy of training or testing data, or model parameters of standard AI models for data analysis, health care, and web browsing. The data exchanged in standard multi-agent robotics tasks like map-building and localizing an agent's position includes odometry, location information, and motion plans; these also demand strong privacy protections. Despite this, there is a relative lack of attention to privacy in robotics from the cryptography, robotics, and policy communities. I see multi-agent robotic systems as a natural candidate for the cryptographic tool of Secure Multi-Party Computation, and pose a research plan that brings privacy to this setting. I have begun initial steps toward building Simultaneous Localization And Mapping (SLAM) for multiple agents while preserving privacy, but building this full pipeline is a sizeable effort in applied cryptography, utilizing not only standard efficient methods and libraries for multi-party computation but also algorithms for secure ranking, stochastic gradient descent and Oblivious RAM.

Verifiability. Much like my work on improving the transparency and verifiability of content moderation, robotic systems (and AI in general) are high-stakes systems that require trust in the specification, and trust in the implementation. I plan to perform a similar policy analysis and cryptographic protocol design process, identifying and implementing specific opportunities for transparency and verifiability in these systems. The desirability of being able to specify a robot or AI's behavior in a cryptographically verifiable manner is especially potent if the current growth of autonomous weapon systems continues—but even in the civilian domain, I hope to make it possible to create regulations and policies for the cryptographic quality assurance of robotic and AI software.

Autonomous systems are one of the greatest tech policy challenges of our time, but cryptography has many tools to offer for improving the authentication, transparency, and privacy of these systems. This is already a challenging topic in the current world of growing virtual AI, but gains tangible physical dangers as we transition toward a more cyber-physical world of self-driving cars and drones. I am excited to be at the forefront of this area, and to use both my policy and cryptography expertise to navigate the challenges of the coming decade.

Cited work

[CSV22] Aloni Cohen, Sarah Scheffler, and Mayank Varia. Can the government compel decryption? don't trust—verify. In *2nd ACM Symposium on Computer Science and Law*, 2022. Author order is alphabetical.

[GHS+21] Yaron Gvili, Julie Ha, Sarah Scheffler, Mayank Varia, Ziling Yang, and Xinyuan Zhang. Turboikos: Improved non-interactive zero knowledge and post-quantum signatures. In *International Conference on Applied Cryptography and Network Security*, pages 365–395. Springer, 2021. Author order is alphabetical.

[GSV21] Yaron Gvili, Sarah Scheffler, and Mayank Varia. Booligero: improved sublinear zero knowledge proofs for boolean circuits. In *International Conference on Financial Cryptography and Data Security*, pages 476–496. Springer, 2021. Author order is alphabetical.

[Scheffler21] Sarah Scheffler. password-ag2pc: An FC-reilient version of EMP-ag2pc., 12 2021. `https://github.com/sarahscheffler/password-ag2pc`.

[SKM23] Sarah Scheffler, Anunay Kulshrestha, and Jonathan Mayer. Public verification for private hash matching, 2023. Forthcoming (conditionally accepted) at IEEE Security & Privacy.

[SM23] Sarah Scheffler and Jonathan Mayer. Systematization of knowledge: Content moderation for end-to-end encryption, 2023. Forthcoming at Privacy Enhancing Technologies Symposium.

[SMK22] Sarah Scheffler, Jonathan Mayer, and Anunay Kulshrestha. Comments on the safety tech challenge fund evaluation criteria, 4 2022. URL: `https://sarahscheffler.net/Comments_on_UK_Safety_Tech_Challenge_Evaluation_Criteria.pdf`.

[SO19] Sarah Scheffler and Jacob Ostling. Dismantling false assumptions about autonomous weapon systems. 2019. Second place in the Student Paper Competition at ACM CSLaw 2019.

[STV22] Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for coyright law's substantial similarity. In *2nd ACM Symposium on Computer Science and Law*, 2022. Author order is alphabetical.

[SV21] Sarah Scheffler and Mayank Varia. Protecting cryptography against compelled self-incrimination. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021. Author order is alphabetical.