This document is v1, and *v2 should be regarded as a different paper*
Take care to cite v1 and link as https://arxiv.org/abs/2012.08347v1

Beyond Privacy Trade-offs with Structured Transparency

Andrew Trask*‡§ Emma Bluemke*†§ Ben Garfinkel‡§
Claudia Ghezzou Cuervas-Mons§ Allan Dafoe‡

Abstract

Many socially valuable activities depend on sensitive information, such as medical research, public health policies, political coordination, and personalized digital services. This is often posed as an inherent privacy trade-off: we can benefit from data analysis or retain data privacy, but not both. Across several disciplines, a vast amount of effort has been directed toward overcoming this trade-off to enable productive uses of information without also enabling undesired misuse, a goal we term 'structured transparency'. In this paper, we provide an overview of the frontier of research seeking to develop structured transparency. We offer a general theoretical framework and vocabulary, including characterizing the fundamental components — input privacy, output privacy, input verification, output verification, and flow governance — and fundamental problems of copying, bundling, and recursive oversight. We argue that these barriers are less fundamental than they often appear. Recent progress in developing 'privacy-enhancing technologies' (PETs), such as secure computation and federated learning, may substantially reduce lingering use-misuse trade-offs in a number of domains. We conclude with several illustrations of structured transparency — in open research, energy management, and credit scoring systems — and a discussion of the risks of misuse of these tools.

1 Introduction

We share medical details with doctors in order to receive appropriate care. We share location data with mapping apps in order to receive relevant directions. We allow tracking of our heartbeat, exercise, and sleep patterns because analysis thereof can improve our well-being. Every day, we grant access to personal information to buy goods, make use of services, and collaborate. But this is not without risk: allowing others to use our data can open the door to misuse, ranging from manipulation and public exposure to theft and discrimination. And even if there are many cases where the benefits still exceed the expected harms, there are some where we – at the individual or societal level – choose not to reap the gains of a potential exchange because the risk of misuse is too great.

The list of problems that could be solved if the right analytical tools (such as machine learning techniques) could gain access to the relevant data (ranging from personal health information to state secrets) is vast; it includes curing diseases, addressing coordination issues and even preventing war. But under the current paradigm of information exchange, actors typically have little incentive to grant access to this data.

*Equal contribution
†Institute of Biomedical Engineering, Department of Engineering Sciences, University of Oxford
‡Centre for the Governance of AI, University of Oxford.
§OpenMined

Preprint. Under review.

Disciplines ranging from computer science to law have dedicated significant thought to the question of how to enable desired uses of information while preventing misuse. In this paper, we use the term 'structured transparency' to describe this aim. Although this moniker and many contemporary techniques it encompasses are new, its objective is ancient. For instance, the practice of casting secret ballots, which dates back at least to classical antiquity, allows citizens to communicate preferences without opening the door to voter intimidation (in theory). More recent examples of systems designed to support structured transparency include bomb-sniffing dogs, which can report that a bag is safe without revealing further detail about its contents, and end-to-end encryption systems, which allow services to deliver messages without being able to read them.

While the aims and methods of structured transparency feature in many disciplines, the subject is seldom discussed in general terms. Researchers in different fields typically develop distinct frameworks and terminology for grappling with these topics. In this paper, we have chosen to highlight the general problem for four main reasons.

First, significant commonalities among different manifestations of the problem suggest great opportunity for the cross-pollination of insights: knowledge well-established in one domain might have as-yet-unrecognised significance for others. For example, researchers studying secrecy-preserving nuclear warhead verification have recently begun to draw inspiration from work in cryptography [26]. Discussing structured transparency in general terms may further catalyze cross-disciplinary knowledge transfer.

Second, we believe that there has never been a more *important* time for research on structured transparency. The rise of digital technology has allowed the collection and analysis of sensitive information at an unprecedented scale. While this has led to many productive uses, ranging from groundbreaking research to helpful mobile apps, it has simultaneously introduced threats to privacy and social stability [52, 61, 47, 56]. Well-grounded concerns about further harm have also slowed the development of valuable applications in high-impact domains. Strategies to reduce such trade-offs would have enormous value.

Third, we believe that there has never been a more *promising* time for research on structured transparency. Recent progress in the development of privacy-enhancing technologies (PETs), such as secure computation and differential privacy techniques, has facilitated levels of structured transparency that would previously have been impractical [40, 14]. Although these technologies have broad relevance, awareness of them has yet to percolate to many fields.

Finally, we believe that taking advantage of these opportunities will require interdisciplinary collaboration. The challenge of designing privacy-preserving surveillance protocols, for example, clearly contains myriad technical and institutional components. No single PET will serve as a panacea for all of the problems at hand, and some PETs, while improving privacy along certain dimensions, could also introduce new problems, such as decreased accountability, that will need to be addressed through other means. Otherwise, systems could become increasingly vulnerable to socially destructive outcomes including money laundering, voter manipulation, or algorithmic discrimination [55, 34, 59, 15]. Effectively advancing the cause of structured transparency will require a patchwork of solutions – some technical, some institutional, some societal – drawn from a wide assortment of disciplines. We hope that a shared perspective on structured transparency can lay the foundation necessary for such work.

In this paper, we describe a general framework for understanding structured transparency. We then discuss the core components of structured transparency and the primary barriers to achieving it, and explain how emerging PETs could help reduce these barriers. Next we provide three present-day use cases that illustrate how these technologies can enforce precise information flows and, as a result, minimize risks of misuse. Lastly, we describe potential directions for future work. As noted above, structured transparency is inherently a multidisciplinary topic, with deep roots in several fields. This paper represents an initial attempt to spark discussion with a broad range of researchers and receive feedback.

2 Structured Transparency and Ideal Information Flows

2.1 Information Flow Desiderata

A system exhibits structured transparency to the extent that it conforms with a set of predetermined standards about who should be able to know what, when they should be able to know it, and what they should able to do with this knowledge; in other words, if it enforces a desired information flow. The standard of structured transparency in a given circumstance thus exists relative to the desiderata of the specific use case. Accordingly, the architect of a system that harnesses structured transparency must answer both a question of ends (what constraints does the ideal flow impose on specific actors) and means (how can that ideal be achieved without any data leakage). For example, the designers of a voting system might set goals such as 'all votes should be counted and the public should know the true vote count, but nobody should know who, aside from themselves, voted for whom' and then deploy available technology, such as secret ballots, to obtain this outcome. In this paper, we focus primarily on how a set of rapidly-developing capabilities can serve as powerful *means* for designing more precise information flows. However, we also believe that it is important to comment briefly on prior work addressing the question of *ends*.

Many academic literatures examine the societal consequences of the flow of information. Within international relations, for instance, a great deal of work deals with the role of private information in preventing peaceful resolutions to conflicts [22, 21]. Economists have also extensively studied how the flow of information within a market can support the efficient pricing of goods and services. And game theorists have shown how information exchange can be essential to escaping certain collective problems [32, 8, 7, 51, 25].

However, the extant work most relevant to this paper is the framework of 'contextual integrity' proposed by Nissenbaum et al. It asserts that what people care most about is not simply restricting access to certain kinds of information but ensuring that it flows appropriately. Societies have developed systems of norms that govern the flow of personal information in distinct social contexts (e.g. education, healthcare, and politics [11]). These norms protect people and groups against harm and balance the distribution of power [45, 46].

In a globalized world where much information is stored digitally in forms that permit easy copying, sale, and exploitation, enforcing these context-relative informational norms becomes increasingly challenging. Violations of privacy therefore proliferate, along with associated materials harms such as theft and discrimination [2, 56]. In the framework of contextual integrity, an ideal information flow is one that would enable us to collaborate over information while ensuring that information is used only for the context-relative 'approved' purposes.

In summary, the question of what constitutes an 'ideal information flow' is complex. Communities with different preferences, values, and ways of thinking will tend to give very different answers. For this reason, the remainder of the paper focuses solely on the question of means: how can one enforce a desired information flow with clearly-defined desiderata?

2.2 Information Flow Taxonomy

Information flows vary significantly in terms of the challenges they pose for structured transparency. Below we describe three broad categories in order of ascending complexity. These offer a sample of the diversity of situations in which one might want to apply structured transparency, and also serve as helpful archetypes that we will reference throughout the paper.

Messaging flows seek to transfer a bit of information from one party to another such that it can be trusted by the receiver (verified) and does not reveal other bits. This type of transfer can face structured transparency challenges in practice, but existing social institutions possess useful techniques for solving many such issues. In the analog world, a sealed envelope enables the postal service to deliver a message without inspecting its contents. In the digital world encrypted messaging performs a similar function. Greater difficulty arises when existing systems do not allow for easy *unbundling* of relevant and irrelevant information (we examine this situation when describing the bundling problem in section 4).

Service provider flows involve an external entity (service provider) performing computation within the flow before returning the results to the data owner. The introduction of computation makes

such flows somewhat more complicated: unfortunately, in nearly every case a trusted third-party must be able to view all of the inputs to the computation in order to generate the correct output. Consider, for example, almost any situation – analog or digital – in which a service requires personalization: from necessities such as doctors' visits to conveniences such as automated film recommendations, customized services nearly always require sharing personal data with the service provider.

Finally, **aggregation flows** contain all of the challenges of service-provider flows with the additional requirement that the aggregating entity itself intends to learn a fact, rather than simply process and forward information. This could be a high-level trend such as average user computing statistics or a needle-in-a-haystack problem that involves pinpointing a specific fact or individual.

3 The Components of Structured Transparency

Structured transparency consists of a number of sub-problems: input and output privacy, input and output verification, and flow governance structures. Not every situation requires that all be explicitly addressed, but most trade-offs can be reduced to some combination of these issues. Below, we describe each of these from the perspective of someone participating in an information flow.

Input privacy refers to the ability to process information that is hidden from you and to allow others to process your information without revealing it to them[5]. Consider the sealed envelope as described above: this allows information (the contents of a letter) to be processed (transmitted) by a mail service without that intermediary gaining access to it.

Output privacy allows you to receive/read the output of an information flow without being able to infer further information about the input and, symmetrically, to contribute to the input of an information flow without worrying that the later output could be reverse engineered to learn about your input [14, 19].

While the aim of output privacy is similar to that of input privacy (both relate to protecting providers of inputs to an information flow), they are not the same. Input privacy is concerned with *facilitating the flow of information* without leaking collateral data, whereas output privacy is concerned with *preventing the output of the flow from being reverse engineered* to reveal additional information about the input or sender. Input privacy is concerned with preventing parties participating in *computation* from learning anything[6], whereas output privacy is concerned with preventing the *recipient* from inferring the inputs. Returning to the sealed letter example, output privacy in this context would guarantee that the letter recipient could not deduce sensitive information that the author did not want to include in the message.

At first glance, satisfying input and output privacy may appear sufficient to enforce any information flow. However, most use cases require balancing these criteria with the informational burdens necessary for recipients to trust the data they receive. Privacy, in other words, exists in tension with verification.

Input verification allows you to verify that information you receive from an information flow is sourced from entities you trust, and (symmetrically), it allows you to send information such that the output can be verifiably associated to you. When you sign a document, you make a mark which (in theory) only you can make – a signal to readers of the document that you approved the information contained therein. Novel input verification techniques empower a signer to verify *specific* attributes of an input to an information flow, such as that it came from a trusted source or that it happened within a specific date range.

Output verification allows you to verify attributes of any information processing (computation) within an information flow. In analog systems that contain little to no computation, the difference between input and output verification can be nuanced or nonexistent. Process auditing by an external party, however, is an example of output verification. If a letter goes missing, one might audit tracking records. Likewise, a tax auditor might verify that the flow of information (in this case, funds) from a company's account to their employees' accounts obeys relevant tax codes.

[5] More formally, input privacy is satisfied when multiple data-holding parties can provide inputs to one or more computing parties, who can provide computation without having the ability to know the values of the respective inputs, intermediate variables, or outputs.

[6] Except the output destined for them (if any)

Finally, **flow governance** is satisfied if each party with concern/standing over how information should be used has guarantees that the information flow will preserve that intended use. Even if a flow satisfies a suitable combination of input and output privacy and input and output verification, questions still remain concerning what authority can *modify* the flow. In real-world scenarios, a wide variety of governance mechanisms exist for this purpose (escrow and executorship stand out as widely-used legal devices). Within the physical world, multi-key safety-deposit boxes for holding secure documents accomplish similar goals.

To recap these traits in the context of sending a letter: input privacy is the envelope that prevents unauthorized access, output privacy is withholding sensitive personal information from the letter itself, input verification is the signature on the letter, output verification is a wax seal proving to the recipient that the letter has not been tampered with, and flow governance is equivalent to the letter being shipped in a safe with a combination lock only the recipient knows.

4 Limitations of Structured Transparency in Practice

Satisfying all of the criteria of structured transparency can prove difficult to achieve in practice. For example, often input and output verification can come at the cost of input and output privacy. We view the core barriers to structured transparency in the analog world as consisting of three key technical problems: the copy problem, the bundling problem, and the recursive oversight problem.

The copy problem lies at the root of many structured transparency issues. When a bit of information is shared, governance over its use is released to the recipient, who is then typically free from any further technical limitation to misuse (though ex-post legal or social sanctions may exist). As a result, when deciding whether to share information, data owners often face a trade-off that pits the benefits of sharing with the risks of misuse.

The bundling problem arises from consequences of the copy problem. It is often difficult to share a bit of information without also needing to reveal additional bits because either the conventional encoding will not allow individual bits to be shared or a bit cannot be trusted/verified without the context of other relevant bits. An example is the use of a surveillance video: many pieces of irrelevant (and potentially invasive) information are shared to contextualize the critical piece(s) of information (e.g. whether a suspect was in a particular location). Another example is a driver's license, which reveals all of the details on the card in order to verify a single piece of information, namely whether the individual is old enough to enter a given venue. Put another way, while cutting out your birthday from your license would allow you to hide the other information on your card, it would not suffice to enter an age-restricted establishment.

The use of third-party oversight institutions can solve issues caused by the copy and bundling problems. In doing so, however, this solution presents a third issue: **the recursive oversight problem**. Direct oversight of information only leads to another, even more well informed entity who could misuse information (i.e., "who watches the watchers?" — it can be hard to hold those in power accountable).

In practice, these three hurdles constrain our ability to satisfy the guarantees of structured transparency in a wide variety of domains. The consequence is a great number of data sharing dilemmas: either useful-but-sensitive data goes unused, or society has to absorb the costs of sharing information with misaligned actors. Without new capabilities that could enable more precisely-controlled information flows, we will remain constrained by a Pareto frontier that places these desiderata in tension with one another.

5 Tools for Technical Structured Transparency

Fortunately, many information flows constrained by the bundling and recursive enforcement problems can be addressed elegantly in the digital domain. And while the copy problem itself can never truly be solved (and arguably should not be [18]), relevant digital tools, when combined with other social and legal measures, could effectively mitigate the harms it poses.

Such technical tools can be loosely grouped into responses to the five sub-problems of structured transparency outlined in Section 3.

5.1 Technical Input Privacy

All proposed within the last half-century, technical input privacy tools come primarily from the field of cryptography — public-key cryptography, end-to-end encryption, secure multi-party computation, homomorphic encryption, functional encryption, garbled-circuits, oblivious RAM, federated learning, on-device analysis, and secure enclaves are several popular (and overlapping) techniques capable of providing input privacy [5, 38, 60, 9, 53, 24, 6, 27, 41, 13, 14, 42, 58]. Many of these techniques can theoretically facilitate any arbitrary computation (also known as 'Turing-complete computation') while keeping the computation inputs secret from all parties involved. These methods differ in terms of performance. Homomorphic encryption requires heavy computation even for relatively simple information flows, while secure multi-party computation requires less computation but greater message volume between the various parties in the flow (increased network overhead) [14]. The field still lacks the general-purpose software implementations necessary for widespread use, but this is an active and quickly-maturing area of research.

The most important implication of technical tools for input privacy is that they can, theoretically, perform a task impossible in the analog world: achieving service provider information flows without a trusted third-party [14]. In other words, technical input privacy tools could allow service providers to process data without being able to see it or use it for other purposes outside of the governed information flow. This means that claims along the lines of "we need a copy of the data in order to provide a service that takes it as input" will loses legitimacy as technical input privacy tools mature.

Aggregation flows also stand to benefit from input privacy tools: under the right configuration, the aggregator will learn only the output intended for them, and will be unable to observe the computation inputs, or any outputs not agreed upon with the input flow providers.

5.2 Technical Output Privacy

Early forms of non-technical output privacy focused on redacting sensitive data-points or threatening reprisal via legal or other means against those who reverse engineer inputs. For highly bundled data, redaction can prove very challenging — consider trying to remove 'personal' information from an MRI image without corrupting information useful to cancer researchers. And even when one can redact information easily (such as removing names and SSNs from a database), the advent of big data has demonstrated that 'data anonymization' techniques can sometimes be reversed if enough information is available [50].

However, technical output privacy tools (chiefly, differential privacy and related techniques) can provide strict upper bounds on the likelihood that a data point could be reverse-engineered [19, 20]. This capability is useful in many settings, but it has particular significance in aggregator flows where the actor processing the information is performing statistical analysis; with differential privacy, aggregator flows can reveal high-level insights without ever observing individuals' data in detail. This holds great promise for preesrving privacy in the context of scientific inquiry and surveillance (such as public-health surveillance used to track the progression of COVID-19).

5.3 Input Verification

Technical input verification tools stand out as perhaps the most promising and least utilized of the technical structured transparency tools. Within the context of message flows, they are robust, performant and proven at scale, but rarely used owing to a lack of awareness or usable engineering tools. Most input verification techniques use some combination of public-key infrastructure (SSI, Key Transparency, etc.), cryptographic signatures, input privacy techniques with active security, and zero-knowledge proofs [28, 23, 39, 10]. These methods can allow an actor to verify a specific attribute such that the information flow output contains cryptographic proof of this verification.

For example, returning to the driver's license example: normally, somebody inspecting a driver's license views the card in its entirety (a date-of-birth removed from the rest of card would carry no weight from a verification perspective). Technical input verification tools do not suffer from this constraint[54, 10]. Attributes can be individually verified and revealed within an information flow. Critically, this allows for high levels of both input privacy and input verification, largely eliminating one of the trade-offs that characterizes attempts to achieve structured transparency in analog settings.

Much like input privacy, however, the greatest promise of input verification techniques lies in the ability to verify the inputs to arbitrary computations. Perhaps the simplest example of this is that of verified voting, wherein an election authority can publish the final vote with cryptographic evidence guaranteeing to each voter that their vote was included in the final tally. Because technical input verification techniques make it possible to verify properties of specific inputs selectively, it enables systems which, in practice, are both more private and more strongly verified than analog equivalents.

5.4 Output Verification

The key limitation of output verification tools in many analog contexts is that the verifier must examine the data in order to perform the verification. A tax auditor must inspect cash flows in detail in order to determine whether fraud has occurred, and a financial regulator must have access to credit score inputs and outputs in order to ascertain whether loans have been distributed fairly. This in turn relates to the recursive oversight problem: effective oversight requires granting access that opens the door to more potential misuse.

However, when combined with the aforementioned input privacy techniques, technical tools for output verification can address this challenge. An external auditor could verify properties of an information flow without learning anything beyond the output of targeted tests (e.g. searching for patterns reflective of fraud) while also ensuring that the tests ran correctly. This has significant implications for increasing the precision, scale, and security of auditing institutions, potentially facilitating new types of checks and balances and fairer distributions of power. It also relates to ongoing research such as that conducted in the fairness, accountability, and transparency in machine learning (FATML) community [16].

5.5 Flow Governance

As noted above, information flow governance in the analog world typically employs legal and physical means. The primary benefit of technical flow governance tools is that their scale and efficiency can greatly exceed that of their analog counterparts (See "policy enforcement" in [14]). Secure multi-party computation (SMPC) represents an excellent example of this: SMPC enables the selection of arbitrary parties to govern the flow over arbitrary information, limited only by the compute and network resources of the chosen information shareholders. Trust in the system need not rely on threats of legal retribution against offenders when hard cryptographic limitations can prohibit forbidden behavior.

Flow governance can also be enacted over computation [14]. Whereas service provider flows and aggregation flows previously required a trusted third party to observe all inputs to a computation in order to create the output, technical flow governance systems such as SMPC have no such limitation. Via computation over encrypted, mutually governed information, digital systems can enforce agreed-upon checks and balances between shareholders in ways that are impossible for physical systems.

5.6 In Combination: A New Frontier for Structured Transparency

Together, these tools help address the question of how to enable desired uses of information while avoiding misuse. They enable much more precise social and technical arrangements than their analog predecessors. Most importantly, they provide (1) the ability to unbundle information such that one needs to share only the bits necessary for a collaboration, and (2) a solution to the recursive enforcement problem such that small actors can audit information to which they do not themselves have access.

Technology alone cannot vanquish all information-sharing issues, but these new capabilities can empower complementary legal and social systems to deliver holistic solutions that would previously have been impossible. Consider a mobile app that collects and stores large volumes of personal information about its users, perhaps as a seemingly-necessary component of supplying whatever service it offers (directions, medical advice, etc.). Absent modern structured transparency tooling, users would simply have to conduct a cost-benefit analysis and accept the risks of misuse as the price of the product. With awareness of such techniques, however, legislators and regulators could mandate that any personal data remain on device, or that any off-device operations performed on such data must be suitably privacy-preserving. Advocacy and activist groups could make similar requests, and

consumers could expect and demand more of the services offered to them. Thus, these technologies, in concert with regulators, ethicists, and relevant governance bodies can enforce adherence to more responsible information flows.

6 Illustrations of structured transparency

Above, we have described the goals and capabilities of structured transparency mainly in the abstract. Below, we turn our attention to a handful of domain-specific use cases that illustrate how, by expanding the privacy-performance Pareto frontier, such technologies can lead to meaningful advances.

6.1 Improved Data Flows for Open Research

For years, the open-data movement in research has expressed concern that the inability to share data safely hampers research progress [44, 12, 49, 37, 1, 30, 17, 31, 3, 4, 29, 43]. The core issue at play here is a textbook example of the copy problem: once an individual or institution shares a copy of their data, it becomes extremely difficult to control what the recipient(s) might do with it. Structured transparency enables information flows that can answer specific research questions while letting the data owners maintain control over the only copy of their personal information. Depending on the level of trust between parties and the nature of the research, combinations of the following techniques could make that workflow possible.

Technical input privacy techniques could enable data owners (hospitals, labs, statistics offices, etc.) to grant researchers the ability to perform specific computations over their data without providing access for any other operations. This allows the researcher to answer their research question ('what is the mean weight of newborns in this city?', 'how does my neural network perform on your patient MRI dataset?') without the owner facing the perils of the copy problem [35]. Additionally, using technical output privacy techniques, data owners could prevent reverse engineering of the computation output to reconstruct patient information.

Input verification techniques could allow data owners to prove to researchers various attributes of the dataset, such as whether or not it was used in another experiment (a potentially important feature for research reproducibility). In situations that feature a competitive relationship between institutions or research groups, output verification may also be required. It could, for example, be used to prove that a key statistical result was actually computed by the data owner using the computations requested by the researcher (as opposed to stemming from shortcuts or mistakes).

Finally, flow governance could distribute control across third parties (e.g. funding bodies, stakeholders in a collaboration network, groups safeguarding rights for vulnerable populations, etc.) to enable especially sensitive to remain available for appropriate research while minimizing risk of misuse.

In summary, stronger, more precise and more automated controls over data sharing could make more scientific data available for research, increasing the pace of scientific research in many empirically-driven fields.

6.2 Large-Scale Collaboration for Social Good

Coordination at scale could alleviate a number of societal challenges. Often, however, privacy concerns impede such arrangements. To take a specific example, consider energy efficiency: the collection of detailed energy usage data from smart meters has immense potential both to reduce gratuitous carbon emissions and to save consumers money. This data, however, could also be used to infer occupancy and activity patterns at a high level of detail, down to which television channel is being watched [57].

Structured transparency tools could enable the use of insights from usage data to optimize energy consumption, while letting customers maintain control over the only copy of their data. As with the researchers in the example above, the meter company does not need a full copy of this data in order to achieve their goals; they only require the output of specific computations (statistics, model training, or model testing).

Technical input privacy techniques could allow providers to perform their smart service (whatever that may be) using energy usage data without ever seeing it in its entirety. Using technical output privacy techniques, smart-meter companies could prevent any reverse engineering that might infringe customer privacy. Flow governance could distribute control beyond the private companies to third parties such as consumer interest groups unlikely to collude against the consumer in a data attack (e.g. environmental protection or privacy rights activist organizations). In this use case, since no party has any incentive to poison the data or computation, input verification and output verification are likely unnecessary. This means that the combination of input and output privacy with flow governance would likely be enough to support this information flow.

6.3 Verification and Bias

Banks' ability to predict the likelihood of a customer defaulting on a loan can play an incredibly important function in society, ideally facilitating the quantification of financial risk and the efficient allocation of resources. However, the data most predictive of default risk includes some highly personal information. Modern credit scoring systems incorporate details ranging from court records and landlord testimony to employer information, insurance histories and, in some cases, social media profiles [33, 36].

Credit scoring agencies have a strong incentive to obtain this information because of the light it can shed on individuals' creditworthiness. And financial institutions often have no obligation to reveal their credit algorithm – or a list of inputs it relies on – to users. This is generally due to arguments that releasing further algorithmic detail could open the algorithms to being 'gamed' by consumers and/or involve surrendering proprietary intellectual property to competitors [48]. In addition to creating information asymmetry and power imbalances, such an opaque situation prompts questions about the fairness and bias: consumers have no reason to trust a process about which they know nothing.

Technical input privacy techniques could allow banks to generate credit scores without needing to centralize massive amounts of personal customer data. Technical input verification techniques could enable the sourcing of information directly from consumers — alleviating the incentive for a back-channel private data marketplace and reducing the opportunity for intentional or unintentional tampering with said data (modifications would invalidate the signature). Perhaps most importantly, output verification techniques could allow external auditors to evaluate the fairness and equity of credit scoring algorithms without ever making the algorithm itself public.

6.4 Potential Misuse of these Tools

It is important to note that the capacity to enforce information flows more precisely by no means guarantees that the flow in question is ethically designed or socially beneficial. A malevolent actor could train an algorithm on users' online activity in a perfectly encrypted way and use differential privacy to achieve input and output privacy, but still use the resulting insights to manipulate users in ways that are not in the their best interest.

All of this highlights the distinction made in Section 2.1 (information flow desiderata) between means and ends. Structured transparency technologies offer an avenue for achieving 'ideal' information flows, but one must always ask, 'ideal for whom'? The ideals of the set of stakeholders empowered to design the system may differ greatly from those of the average system user. In the examples we detailed above, we advocate for a certain set of ends, but the desirability of these is far from self-evident. In short, like many technologies, structured transparency tools are morally agnostic. Whether this toolkit truly results in socially beneficial outcomes depends on a much broader set of considerations than those discussed in this paper.

7 Conclusion

This paper attempts to begin conversations around structured transparency. We have outlined a framework that describes how a set of rapidly-developing technical tools can help design and enforce far more precise information flows than those currently in widespread use. These methods are not complete solutions in and of themselves, but they enable selective sharing and disclosure of information and thereby offer an opportunity to expand the Pareto frontier that has historically defined trade-offs between privacy and performance. We have set forth several examples illustrating

how structured transparency tools could enable the use of sensitive information for social good while minimizing the risk of misuse. These tools sit at the intersection of a number of active fields of research and have potential applications for many more. It is our hope that this framework acts as a useful bridge between disciplines, and we look forward to receiving feedback.

8 Acknowledgements

This work is the culmination of over two years of both formal research and casual conversation between a wide community of researchers within the AI Safety and privacy communities around the Centre for the Governance of AI and OpenMined. In particular, we'd like to thank and acknowledge Teddy Collins for deep and extensive conversation, idea refinement, and deep editing of the paper. We'd like to thank and acknowledge William Isaac for continuous feedback, brainstorming, and ideation around the paper over the last two years as well as Iason Gabriel for shepherding the narratives which were chosen for the paper - providing extensive feedback and advice on paper scope and deep editing suggestions throughout the writing process.

We'd also like to thank the entire GovAI team, many of whom have at some point engaged with ideas now present in this paper. We'd also like to thank the Future of Humanity Institute for their guidance on the project. We'd like to especially thank Eric Drexler for proposing the name "Structured Transparency" as well as for his willingness to give feedback on early iterations of its formalization. Finally, we'd also like to thank Georgios Kaissis, Jan Leike, Vishal Maini, Kenneth Cuckier, Helen Nissenbaum, Phil Blunsom, and Chris Dyer for feedback on ideas and/or early drafts of this work. And finally we'd like to thank the OpenMined community for providing an environment wherein these social and technical ideas could come together.

References

[1] Amr Ali-Eldin, Anneke Zuiderwijk, and Marijn Janssen. A privacy risk assessment model for open data. *Lecture Notes in Business Information Processing*, pages 186–201, 2018.

[2] Julia Angwin. Privacy tools: Opting out from data brokers, 2020.

[3] Giorgio A Ascoli. Sharing neuron data: carrots, sticks, and digital records. *PLoS Biol*, 13(10):e1002275, 2015.

[4] Giorgio A Ascoli, Patricia Maraver, Sumit Nanda, Sridevi Polavaram, and Rubén Armañanzas. Win–win data sharing in neuroscience. *Nature methods*, 14(2):112–116, 2017.

[5] Dan Bogdanov, Peeter Laud, Sven Laur, and Pille Pullonen. From input private to universally composable secure multi-party computation primitives. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 184–198. IEEE, 2014.

[6] Dan Boneh, Amit Sahai, and Brent Waters. Functional encryption: Definitions and challenges. In *Theory of Cryptography Conference*, pages 253–273. Springer, 2011.

[7] Valerio Capraro, Jillian J. Jordan, and David G. Rand. Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments. *Scientific Reports*, 4:6790, 10 2014.

[8] Valerio Capraro and Matjaž Perc. Grand challenges in social physics: In pursuit of moral behavior. *Frontiers in Physics*, 6, 10 2018.

[9] Melissa Chase, Trevor Perrin, and Greg Zaverucha. The signal private group system and anonymous credentials supporting efficient verifiable encryption. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1445–1459, 2020.

[10] David Chaum. Security without identification: Transaction systems to make big brother obsolete. *Commun. ACM*, 28(10):1030–1044, October 1985.

[11] Joshua Cohen and Michael Walzer. Spheres of justice: A defense of pluralism and equality. *The Journal of Philosophy*, 83:457, 08 1986.

[12] European Commission. Facts and figures for open research data.

[13] Victor Costan and Srinivas Devadas. Intel sgx explained. *IACR Cryptol. ePrint Arch.*, 2016(86):1–118, 2016.

[14] Mark Craddock, David Archer, Dan Bogdanov, Adria Gascon, Borja Balle, Kim Laine, Andrew Trask, Mariana Raykova, Matjaz Jug, robert McLellan, Ronald Jansen, Olga Ohrimenko, Simon Wardly, Kristin Lauter, Nigel Smart, Aalekh Sharan, Ira Saxena, Rebecca Wright, Eddie Garcia, and Andy Wall. *https://docs.google.com/document/d/1GYu6UJI81jR8LgooXVDsYk1s6FlM-SbOvo3oLHglFhY/editheading=h.k9fcam5cpoh7*. 2018.

[15] R. Dang-awan, J. A. Piscos, and R. B. Chua. Using sharemind as a tool to develop an internet voting system with secure multiparty computation, 07 2018.

[16] N Diakopoulos, S Friedler, M Arenas, S Barocas, M Hay, B Howe, H Jagadish, K Unsworth, A Sahuguet, S Venkatasubramanian, et al. Principles for accountable algorithms and a social impact statement for algorithms. online: https://www. fatml. org/resources/principles-for-accountable-algorithms, 2017.

[17] Daniel Dietrich, Jonathan Gray, Tim McNamara, Antti Poikola, Rufus Pollock, Julian Tait, and Ton Zijlstra. The open data handbook, 2020.

[18] Cory Doctorow. https://www.youtube.com/watch?v=gbyxbjofgei, 2020.

[19] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[20] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[21] James D Fearon. Rationalist explanations for war. *International organization*, 49(3):379–414, 1995.

[22] James D. Fearon and David D. Laitin. Ethnicity, insurgency, and civil war. *The American Political Science Review*, 97:75–90, 2003.

[23] Uriel Feige, Amos Fiat, and Adi Shamir. Zero-knowledge proofs of identity. *Journal of cryptology*, 1(2):77–94, 1988.

[24] Craig Gentry and Dan Boneh. *A fully homomorphic encryption scheme*, volume 20. Stanford university Stanford, 2009.

[25] Ayman Ghoneim, Hussein Abbass, and Michael Barlow. Information sharing in the iterated prisoner's dilemma game, 2007.

[26] Alexander Glaser, Boaz Barak, and Robert J. Goldston. A zero-knowledge protocol for nuclear warhead verification. *Nature*, 510:497–502, 06 2014.

[27] Oded Goldreich. Towards a theory of software protection and simulation by oblivious rams. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 182–194, 1987.

[28] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on computing*, 18(1):186–208, 1989.

[29] Julie Gould. Data sharing: Why it doesn't happen. *Nature Jobs*, 2015.

[30] Ben Green, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan P. Crawford. Open data privacy, 02 2017.

[31] Tess Grynoch. Resource guides: Research data management resources: Open and publicly available data, 2020.

[32] Friedrich Hayek. The use of knowledge in society. *The American Economic Review*, 35(4):519–530, 09 1945.

[33] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33:847–856, 11 2007.

[34] Nesrine Kaaniche, Maryline Laurent, and Sana Belguith. Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey. *Journal of Network and Computer Applications*, 171:102807, 12 2020.

[35] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2:305–311, 6 2020.

[36] E. Kambal, I. Osman, M. Taha, N. Mohammed, and S. Mohammed. Credit scoring using data mining techniques with particular reference to sudanese banks, 08 2013.

[37] Shinji Kobayashi, Thomas Kane, and Chris Paton. The privacy and security implications of open data in healthcare. *Yearbook of Medical Informatics*, 27:041–047, 04 2018.

[38] Ben Laurie. Certificate transparency. *Communications of the ACM*, 57(10):40–46, 2014.

[39] Jake Loftus and Nigel P Smart. Secure outsourced computation. In *International Conference on Cryptology in Africa*, pages 1–20. Springer, 2011.

[40] Natasha McCarthy and Franck Fourniol. The role of technology in governance: The example of privacy enhancing technologies. *Data Policy*, 2, 06 2020.

[41] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[42] Marcela Melara, Aaron Blankstein, Princeton University, Joseph Bonneau, Edward Felten, and Michael Freedman. Open access to the proceedings of the 24th usenix security symposium is sponsored by usenix coniks: Bringing key transparency to end users coniks: Bringing key transparency to end users, 08 2015.

[43] Gina Neff. Why big data won't cure us. *Big data*, 1(3):117–123, 2013.

[44] The Open Data Research Network. The open data research network | open data research network, 2020.

[45] Helen Nissenbaum. Privacy as contextual integrity. *WASHINGTON LAW REVIEW*, 2004.

[46] Helen Nissenbaum, Sebastian Benthall, Anupam Datta, Michael Carl Tschantz, and Piot Mardziel. Origin privacy -protecting privacy in the big-data era new york university approved for public release; distribution unlimited air force materiel command, 2018.

[47] Cathy O'Neil. Life in the age of the algorithm. *Science*, 355:137–137, 01 2017.

[48] Frank Pasquale. *The Black Box Society*. Harvard University Press, 2015.

[49] European Data Portal. Open data and privacy 1 analytical report 3: Open data and privacy analytical report n3. *Luxembourg: Publications Office of the European Union, 2020 © European Union, 2020*, 2020.

[50] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 07 2019.

[51] Francisco C. Santos, Jorge M. Pacheco, and Brian Skyrms. Co-evolution of pre-play signaling and cooperation. *Journal of Theoretical Biology*, 274:30–35, 04 2011.

[52] Bruce Schneier. *Data and Goliath : the hidden battles to collect your data and control your world*. W.W. Norton Company, 2015.

[53] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[54] Sovrin. The sovrin network and zero knowledge proofs. *sovrin.org*.

[55] William Stadler. Risks of privacy-enhancing technologies: Complexity and implications of differential privacy in the context of cybercrime. *Security and Privacy From a Legal, Ethical, and Technical Perspective*, 09 2020.

[56] Carissa Veliz. *Privacy is power : why and how you should take back control of your data*. Transworld Digital, 2020.

[57] Carissa Véliz and Philipp Grunewald. Protecting data privacy is key to a smart energy future. *Nature Energy*, 3:702–704, 07 2018.

[58] Fennie Wang and Primavera De Filippi. Self-sovereign identity in a globalized world: Credentials-based identity systems as a driver for economic inclusion. *Frontiers in Blockchain*, 2, 01 2020.

[59] Jan Willemson. How not to use a privacy-preserving computation platform: Case study of a voting application. *Computer Security*, 11980:111–121, 2020.

[60] A. Yao. Protocols for secure computations. In *FOCS 1982*, 1982.

[61] Shoshana Zuboff. *The age of surveillance capitalism : the fight for a human future at the new frontier of power*. Publicaffairs, 2019.

9 Appendix

9.1 Appendix A: Abilities and Limitations of Technical Structured Transparency

Tools for technical structured transparency move the ball forward in capability over previous analog approaches (such as those found in the social and legal domains). In particular, service-provider and aggregation flows previously required a trusted-third party to view all of the inputs to an information flow and generate the output. Tools for technical structured transparency endow us with a better "black box" for our information flows which, more than simply transferring information (like an envelope), can actually perform computations without any party seeing the inputs (and therefore opening them up to the risks of the copy problem).

However, looking forward it becomes clear that even if tools for technical structured transparency were to become fully realized throughout society, there are still limitations. Specifically, not every information flow within the real world can be performed entirely within the digital domain. Sometimes large systems cannot entirely be put into a single "black box" to hide intermediate variables, and several smaller information flows have to pass information to and from various parties throughout the larger system.

While additional limitations may exist, the primary paradigm we observe is that technical structured transparency tools can theoretically provide arbitrary information flows and arbitrary information governance up until a party in the real world needs to take a physical action. Related to the branching problem in computer science, physical action in the real world leaks information in a way which cannot be captured without simply not taking the action, and non-technical structured transparency tools (legal and social) must take over to govern this analog information.

Thus, some information flows within society can benefit greatly from technical structured transparency tools, and others less so. We exposit this through three examples showing the full range of this dynamic. First, consider a taxi driver receiving your home address as a pickup location. In this case, the taxi driver must take actions in the physical world (turning the steering wheel and pressing pedals) in order to provide the service, and thus technical structured transparency is of little help. On the flip side, a software program for machine translation, hosted in the cloud, could allow a customer to encrypt their document, send it to the cloud for translation, and return the encrypted result back

to the customer without the cloud provider ever having the ability to read (either translation of) the document they processed. Because no physical action in the physical world is required, technical structured transparency provides a strong improvement over the information flow of a similar process in the analog world (hiring a translator).

While this might seem obvious at first, the nuance occurs when designing systems with a mixture of digital processing and physical action. For example, one commonly discussed use-case of smartphone applications during the COVID pandemic is to deny COVID-positive individuals access to particularly sensitive areas (such as nursing homes). However, in the early days of the pandemic, on several occasions individuals were physically assaulted because they were believed to be a potential COVID carrier. Thus, the concern arose that if someone was turned away from an establishment, everyone in their immediate surroundings would know that they were a COVID patient and potentially harm them. It might appear that technical structured transparency is of no use in this situation, individuals who are sick need to be denied entrance.

However, while technical structured transparency tools cannot perfectly alleviate the situation, technical output privacy techniques can be applied to the output of the information flow indicating who is likely COVID positive. The result is that 90% of the time an individual is denied entrance, it is a false positive and the individual is instructed to walk back to their car and re-run the app from within their car. This gives any person denied access plausible deniability, as everyone else in line to enter the secure facility knows that 90% of the time people turned away aren't actually sick. Note that we do not give this illustration as an endorsement of apps for this use case, but merely to illustrate a real-life scenario in which technical structured transparency serves only a partial role because actions themselves leak information in the real world.