JOURNAL
OF
LAW & INNOVATION

VOL. 2 OCTOBER 2019 NO. 1

HTTPS://WWW.LAW.UPENN.EDU/INSTITUTES/CTIC/JLI/

JOURNAL
OF
LAW & INNOVATION

VOL. 2 OCTOBER 2019 NO. 1

CONTENTS

ARTICLES

Board of Officers
Volume 2

JOURNAL
OF
LAW & INNOVATION

VOL. 2 OCTOBER 2019 NO. 1

ARTICLE

ALL SMART CONTRACTS ARE AMBIGUOUS

JAMES GRIMMELMANN†

Smart contracts are written in programming languages rather than in natural languages. This might seem to insulate them from ambiguity, because the meaning of a program is determined by technical facts rather than by social ones. It does not. Smart contracts can be ambiguous, too, because technical facts depend on socially determined ones. To give meaning to a computer program, a community of programmers and users must agree on the semantics of the programming language in which it is written. This is a social process, and a review of some famous controversies involving blockchains and smart contracts shows that it regularly creates serious ambiguities. In the most famous case, The DAO hack, more than $150 million in virtual currency turned on the

---

[†] Professor of Law, Cornell Tech and Cornell Law School. I presented earlier versions of this essay at the Algorithms, Big Data, and Contracting symposium at the University of Pennsylvania Law School, at the Princeton Center for Information and Technology Policy, and to a group of students at Cornell Tech. My thanks to the participants, and to Aislinn Black, Andrew Appel, Matthew D'Amore, Karen Levy, Stephen Sachs, and Lawrence Solum. This essay may be freely reused under the terms of the Creative Commons Attribution 4.0 International license, https://creativecommons.org/licenses/by/4.0.

contested semantics of a blockchain-based smart-contract programming language.

> *Those who lack intimacy with the machine cannot be expected a priori to have insight into its limitations. . . . Even in the most formal and most mechanical of domains, trust in the machine cannot entirely replace trust in the human collectivity.*[1]

INTRODUCTION

"Smart contracts" are neither smart nor contracts,[2] but the name has stuck. Instead, they are mechanisms that enforce agreements using software rather than law.[3] The contracting parties write a computer program that embodies their agreement. The program updates as they perform their obligations, and automatically delivers the appropriate resources to them as they become entitled to payment. Smart contracts

---

[1] DONALD A. MACKENZIE, MECHANIZING PROOF: COMPUTING, RISK, AND TRUST 334 (2001).

[2] Ed Felten, *Smart Contracts: Neither Smart nor Contracts?*, FREEDOM TO TINKER (Feb. 20, 2017), https://freedom-to-tinker.com/2017/02/20/smart-contracts-neither-smart-not-contracts. I will refer to "smart contracts" and "legal contracts" in this essay.

[3] *Id*. For more on the terminology and a discussion of how smart contracts relate to legal contracts, see *infra* Part I.

range from simple escrow schemes to immensely complicated joint ventures.

One argument in favor of smart contracts emphasizes the clarity and certainty of code. Legal contracts are written in natural language, which is full of ambiguity, and must be interpreted subjectively by fallible humans. Smart contracts are written in programming languages, which are unambiguous and executed objectively by infallible computers. The result is that anyone reading a smart contract can predict what it will do in response to any conceivable set of events. Legal contracts are ambiguous; smart contracts are not.

So goes the argument. But it is wrong. Smart contracts do not eliminate ambiguity — they hide it. The meaning of a legal contract is a social fact. So too is the meaning of a smart contract. It does not depend directly on what people think it means when they read it, as a legal contract's meaning does. Instead, it depends indirectly on what people think about the computer systems on which it runs. Smart contracts may in fact be more predictable and consistent than legal contracts. Or they may not. But the argument that smart contracts are not ambiguous because they cannot be is false. Worse than that, it is dangerous, because it distracts attention from the hard work required to make smart contracts work in the real world.

Part I of this essay reviews how smart contracts on blockchains work. Part II discusses ambiguity in natural and programming languages. Part III gives examples of ambiguous smart contracts. A brief conclusion then draws out some implications for blockchain governance.

I. SMART CONTRACTS

The defining feature of smart contracts is automation.[4] They are executed by hardware and software — physical and digital systems embedded in the world — rather than by human instructions. Thus, they provide a way for parties to enjoy the benefits of binding contracts without relying on a legal system: private law without a public authority.

The relationship between smart contracts and legal contracts is complicated.[5] It is helpful to make two additional distinctions. One is

---

[4] *See* Nick Szabo, *Smart Contracts: Building Blocks for Digital Markets*, 16 EXTROPY 1, 1 (1996); Nick Szabo, *Formalizing and Securing Relationships on Public Networks*, 2 FIRST MONDAY (1997).

[5] *See generally* J.G. Allen, *Wrapped and Stacked: 'Smart Contracts' and the Interaction of Natural and Formal Language*, 14 EUR. REV. CONT. L. 307 (2018) (explaining the intersections and differences between smart contracts and legal contracts); Lauren Henry Scholz, *Algorithmic*

between *relations of obligation*, like the legal obligation to pay $5 on Tuesday, and the *instruments* which evidence and establish those relations, like an IOU saying, "I will gladly pay you Tuesday for a hamburger today."[6] The other is between *natural* and *formal* languages. Natural languages are used by people to communicate with each other. They can evolve entirely without conscious direction, like English and Mandarin, or they can be created, like Klingon and Esperanto. Formal languages include programming languages, which consist of commands to a computer, as well as various mathematical and logical formalisms.[7]

The paradigm of a legal contract is a relation of legal obligation based on a natural-language instrument. Because legal contracts can be oral or illusory, there can be legal obligations without a corresponding instrument, and vice-versa. In additional, legal contracts can incorporate terms in formal languages. For example, the price term in a contract could be expressed using an algebraic equation or based on the output of a program. The parties' obligations would then be determined in part by the result of a computation.

Obligations can also be *technical* rather than *legal*. A technical obligation is one that is enforced immediately by a system that prevents the prohibited conduct *ex ante* rather than punishing it *ex post*.[8] All but the simplest technical obligations must be based on an instrument, and that instrument must be written in a programming language — this is just another way of saying that computers do only what they are programmed to do. The paradigm of a smart contract is thus a technical obligation based on a formal-language instrument. This is where the conflation of obligation and instrument in smart contracts comes from — and also where it breaks down. Because a legal obligation can be embodied in part in a formal-language instrument, a legal obligation may therefore "wrap" a technical obligation.[9] On the other hand, parties who enter into a technical obligation at the same time may or may not enter into legal obligations effectively wrapping it — or they may even enter into legal obligations without knowing it or intending to.[10] Much

---

Contracts, 20 STAN. TECH. L. REV. 128, 128, 136 (2017) (outlining various uses of smart contracts and arguing that 'black box' algorithmic contracts are likely unenforceable); Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629, 688-89 (2012) (explaining that firms are driven to adopt smart contracts in part because of the advantages of applying computers' high processing power to contractual obligations).

[6] Allen, *supra* note 5.

[7] *Id.*

[8] James Grimmelmann, Note, *Regulation by Software*, 114 YALE L.J. 1719, 1729–30 (2005).

[9] Allen, *supra* note 5.

[10] *See* Adam J. Kolber, *Not-So-Smart Blockchain Contracts and Artificial Responsibility*, 21 STAN. TECH. L. REV. 198, 214-26 (2018) (distinguishing the "code" from the "contract").

of the literature about whether "smart contracts" are "contracts" deals with this last question, but focusing too much on it obscures the other similarities and differences in the analogy.[11]

A. *Smart Contracts*

The turn to automation is motivated by three well-known difficulties with natural language and human institutions. The first is ambiguity — the fear that because legal contracts are written in natural language, they will be interpreted differently by different parties and judges.[12] The second is corruption — the fear that human judges who interpret and enforce legal contracts can be threatened or bribed.[13] A third is enforcement — the fear that parties might be able to ignore a legal judgment by fleeing the jurisdiction, delay, physical force, hiding assets, or never having assets in the first place.[14] These are opportunities for smart contracts to improve on legal contracts; they are also challenges that smart contracts must confront. In this essay, I will focus on ambiguity, although, as we will see, the three are closely related.

Smart contracts are designed to respond to all three of these concerns by expressing contractual terms in a programming language rather than in a natural language.[15] Consider a standard example of a smart contract: a

---

[11] In this essay, I focus on the parallel with contracts, rather than with other kinds of legal instruments, such as wills, statutes, and terms of service, which raise distinct interpretive issues. I even avoid dealing with many interesting legal interpretive issues raised by smart contracts, such as whether they should be regarded as contracts of adhesion.

[12] Max Raskin, *The Law and Legality of Smart Contracts*, 1 GEO. L. TECH. REV. 305, 324-25 (2017). *See also* Surden *supra* note 5; AARON WRIGHT & PRIMAVERA DE FILIPPI, BLOCKCHAIN AND THE LAW: THE RULE OF CODE (2018); Usha Rodrigues, *Law and the Blockchain*, 104 IOWA L. REV. 679, 682 (2019). Or, to quote from Roger Traynor's famous opinion in *Pacific Gas & Electric Co. v. G.W .Thomas Drayage & Rigging Co.*:

> If words had absolute and constant referents, it might be possible to discover contractual intention in the words themselves and in the manner in which they were arranged. Words … do not have absolute and constant referents. A word is a symbol of thought but has no arbitrary and fixed meaning like a symbol of algebra or chemistry. The meaning of particular words or groups of words varies with … context and surrounding circumstances … A word has no meaning apart from these factors; much less does it have an objective meaning, one true meaning.

Pac. Gas & Elec. Co. v. G.W. Thomas Drayage & Rigging Co., 442 P.2d 641, 644 (1968) (citations omitted). A term in a programming language, on the other hand, appears more like a "symbol of algebra" with an "absolute and constant referent." Punch line: symbols of algebra don't have absolute and constant referents, either.

[13] Raskin, *supra* note 12, at 319.

[14] *See* Szabo, *Formalizing and Securing*, *supra* note 4; Szabo, *Building Blocks*, *supra* note 4.

[15] In theory, a smart contract could be implemented in hardware rather than in software. But any hardware sophisticated enough to implement a nontrivial smart contract would need to be

vending machine.[16] Expressing the contract for the sale of a pack of Skittles in a programming language resolves all sources of ambiguity, because programming languages are unambiguous. The machine's code to dispense an item from row C4 when the buyer has inserted $1.50 is completely specified. Committing the contract to the software resolves the fear of corruption, because computers are incorruptible. Threats and offers of bribes literally mean nothing to the vending machine. And the smart contract resolves the concern about enforcement because it takes direct control of the relevant resources. No money, no Skittles.

The vending machine is obviously limited. Scaling it up to a true smart contract platform requires identifying and overcoming its major shortcomings:

1. First, the vending machine is special-purpose: it is good only for spot candy sales. A better smart-contract platform would be general-purpose, capable of being used by many parties for many kinds of contracts.

2. Second, the vending machine's code is unobservable by the user. Unambiguous code can still be malicious. Every time you put a coin in one, you are trusting that its code really does instruct it to dispense Skittles when you push C4. A better smart-contract platform would make contract code visible to affected parties.

3. Third, while the machine is by definition incorruptible, its programmer and its operator are not. You won't get anywhere pointing a gun at a vending machine, but you might if you point a gun at the technician with a key to the coinbox when he comes to restock the Skittles. A better smart-contract platform would be decentralized. The power to supervise and control the execution of the smart-contract code would be dispersed over a large population, so that no individual or small group's corruption threatens the contract.

4. Fourth, the machine is physically vulnerable. If you punch a hole in the window, you can grab all the Skittles you want. A better smart-contract platform would have direct control over resources

---

specified in some way, and that specification is effectively equivalent to a computer program. It is simply a program that is compiled into special-purpose hardware, rather than into object code for execution on general-purpose hardware.

[16] Szabo, *Formalizing and Securing*, *supra* note 4.

> whenever possible. That is, whenever it could it would use virtual resources rather than physical ones.

All of these design goals point in the same direction: put the smart contract on a *blockchain*.

B. *Blockchains*

At heart, a blockchain is a ledger of transactions. It organizes digital records of transactions into discrete chunks (*blocks*), and then maintains a chronological list of those blocks (the *chain*). A chain of blocks: a blockchain. Although the basic computer-science ideas are older,[17] "Satoshi Nakamoto's" Bitcoin proposal put them together in a clever way, greater than the sum of its parts.[18]

The first important design choice is that the transactions in a blockchain are *cryptographically secure*. New transactions are processed only if they are digitally signed by the relevant party (usually the one who pays for them or transfers assets) using a private key that only they (should) know.[19] New transactions are also required to be consistent with the history of transactions on the blockchain: you can't transfer Bitcoin unless you received it in a previous transaction. Together, these consistency constraints mean that only parties who have digital assets are able to use them in transactions.

The second important design choice is that the blockchain is a *distributed* ledger. Every participant has (or could have, if they wished) a complete copy of the entire blockchain. No participant's copy is canonical; all are equally authoritative. Thus, there is no centralized recordkeeper with authority over the ledger. This is where blockchains achieve their resistance to corruption: anyone hoping to tamper with the ledger will need to suborn a significant fraction of participants, not just one.[20]

---

[17] *See* Arvind Narayanan & Jeremy Clark, *Bitcoin's Academic Pedigree*, 60 COMM. OF THE ACM 36 (2017) (describing the basic structure of blockchain). *See generally* FINN BRUNTON, DIGITAL CASH: THE UNKNOWN HISTORY OF THE ANARCHISTS, UTOPIANS, AND TECHNOLOGISTS WHO BUILT CRYPTOCURRENCY (2019).

[18] Satoshi Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System* (2008), https://bitcoin.org/bitcoin.pdf. *See generally* ARVIND NARAYANAN ET AL., BITCOIN AND CRYPTOCURRENCY TECHNOLOGIES: A COMPREHENSIVE INTRODUCTION (2016).

[19] This is possible because with modern public-key encryption, other participants can verify that a message was properly signed by the private key holder even though they do not themselves have the private key.

[20] The redundancy also means that blockchains are practically impervious to hardware errors: any idiosyncratic faulty execution on one participant's computer will be massively outvoted by the collectivity of participants whose computers did not malfunction. Thus, in what follows, I will ignore the philosophical objection that no program can guarantee that it runs correctly on actual hardware. *See, e.g.*, James H. Fetzer, *Program Verification: The Very Idea*, 37

The third important design choice solves a problem introduced by the second. Distributed systems need to reach some form of consensus: if multiple parties can each have copies of the ledger, there must be some way to keep their copies in sync, or to deal with the disagreement if they are not. Bitcoin's mechanism to do so — the Bitcoin *consensus protocol* — is the most ingenious part of Nakamoto's design for Bitcoin and is in some ways the most interesting and influential thing about it.

In brief, the Bitcoin consensus protocol asks participants (called "miners") to accept any valid new block of transactions that one miner proposes — but it makes the process of generating a valid new block onerous and unpredictable. (The difficulty is regularly adjusted so that the entire community of miners can on average generate only one new block every ten minutes.) When a miner broadcasts the block to other miners, they examine it, confirm that it satisfies the consistency constraints, and then with majority approval, add it to the current blockchain. Then the process begins anew to generate the next block.

Incentives are needed to make miners generate and accept blocks. A miner receives a "block reward" of new Bitcoin for each block they successfully generate, and "transaction fees" paid by users to add their transactions to a block. Their incentives to accept valid blocks proposed by other miners come from the value of consensus itself: new blocks can only be added to what everyone else agrees is the current end of the chain. So a miner who fails to approve a valid block may be cutting herself off from future mining rewards: any blocks she generates will not be at the end of the chain. In equilibrium, the dominant individual strategy for individual miners is typically to accept any valid new block and immediately start trying to generate a block that follows it.

C. *Smart Contracts on Blockchains*

Now let us consider how to put smart contracts on a blockchain. The basic idea is simple. There is still a ledger of transactions, maintained in the same way as the Bitcoin blockchain. The difference is that these transactions are richer: they can create and execute computer programs, not just transfer resources.

These programs run on a *virtual machine*. As the name implies, it executes instructions like an actual computer, but it is entirely simulated. The Ethereum blockchain, for example, implements the Ethereum Virtual

---

COMMS. OF THE ACM 1048, 1059–60 (1988). When it comes to hardware faults, the objection is ontologically impeccable but practically irrelevant in this context. The more relevant objection, as I argue below, is that no program can guarantee that *people* will run it as intended.

Machine (EVM). One kind of transaction on the Ethereum blockchain simply transfers its native currency unit — called "Ether" — from one user to another. Another kind of transaction takes a program written in the EVM's native language ("EVM bytecode") and runs it on the EVM.

This last sentence is deceptively simple, so it is worth unpacking. The EVM is a simulated computer. It functions according to rules described in the Ethereum protocol[21] — that is, each participant on the blockchain independently applies those rules to each new transaction and confirms that they yield the same result. The consensus protocol ensures that each user observes the same transfer and program transactions. Thus, just as the participants agree on each user's current balance of Ether because they agree on how each transfer transaction changes those balances, they agree on the EVM's current state because they agree on how each program transaction changes the EVM. The rules are significantly more complicated (though far less complicated than the circuits in a typical physical computer), but they are deterministic.

There are a few more details worth noting. First, EVM bytecode includes instructions for programs to send and receive Ether. A program can transact with users (or with other programs) by executing these instructions. Second, programs can be persistent: one user can load a program into the EVM with an initial transaction, and other users can then interact with it in subsequent transactions (if and how its code allows, that is). Together, these features enable smart contracts: I can offer you a smart contract by loading its terms into the EVM, and you can accept by sending it an appropriate transaction. Third, these program transactions are not free. Ethereum has a complicated metering scheme in which programs consume a resource called "gas" as they run: users must pay (with Ether) for enough gas for the programs they run. The design is both clever and ambitious.

II. AMBIGUITY

It might be hoped that this approach to putting smart contracts on a blockchain solves the three problems with legal contracts identified above. The smart contracts are unambiguous because they are written in programming languages. The smart contracts are incorruptible because control of the blockchain is widely distributed. And enforcement is

---

[21] *See generally* GAVIN WOOD, ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER (2014); ANDREAS M. ANTONOPOULOS & GAVIN WOOD, MASTERING ETHEREUM: BUILDING SMART CONTRACTS AND DAPPS (1st ed. 2018).

automatic because the smart contract directly controls resources on the blockchain. I believe these hopes are overstated.

A. *Ambiguity in Natural Languages*

Consider a famous example of the ambiguity of natural language. In *Frigaliment Importing Co. v. BNS International Sales Corp.*, the parties disagreed on the meaning of "chicken."[22] Their contract called for the delivery of 100,000 pounds of ""US Fresh Frozen Chicken, Grade A, Government Inspected, Eviscerated." The buyer thought that "chicken" meant "a young chicken, suitable for broiling and frying."[23] The seller thought it meant "any bird of that genus."[24] The court considered dictionary definitions, the text of the contract, the parties' negotiations (in a mixture of English and German), evidence of trade usage in the chicken-evisceration industry, USDA inspection standards, and prevailing market prices, only to conclude that there was evidence on both sides, so the plaintiff had failed to carry its burden of "showing that 'chicken' was used in the narrower rather than in the broader sense."[25] In short, "chicken" was ambiguous.

The parties in *Frigaliment* could have prevented their particular dispute if they had written "young chicken suitable for broiling."[26] But that would just have raised further ambiguities in other cases. What counts as "suitable for broiling?" Suitable for broiling at 500 degrees Fahrenheit? 550? For how long? Ambiguity always remains.

The problem is inherent in the nature of natural language, because natural language is inherently social. The meaning of a text is not the (single) meaning its author intended, but the (possibly different and possibly plural) meanings it has within the relevant linguistic community. Even the meanings given in "objective" sources like dictionaries — putting aside all of the interpretive problems of how to read those sources — depend on how people actually use words. And since the legal effect of a contract is determined by the interpretation of its terms, the meaning of a contract is irreducibly social.

---

[22] Frigaliment Importing Co., Ltd., v. BNS Intl Sales Corp., 190 F. Supp. 116, 117 (S.D.N.Y. 1960).

[23] *Id.*

[24] *Id.*

[25] *Id.* at 121.

[26] Or "any bird of the genus *gallus gallus domesticus*" if they had settled on the seller's preferred meaning rather than the buyer's.

B. *Ambiguity in Programming Languages*

To repeat, the meaning of "chicken" is a socially contingent fact. It depends on how people actually use the word in the world. Its meaning can vary and be misunderstood.

It might be argued, however, that the meaning of an expression in a programming language is a technical fact rather than a socially contingent fact. `2**3` in Python will always evaluate to `8`. Its meaning never changes, and if you think it means `9` you are wrong. Meanings that depend on socially contingent facts can be ambiguous, but meanings that depend on technical facts cannot.

This account is wrong. It is true that competent programmers in a given language will agree on a program's meaning (at least for simple programs). And their agreement does depend on technical facts about the language that are independent of particular programmers' idiosyncratic beliefs. But these technical facts are still social, just at a deeper level.

In a nutshell, no computer program can determine its own semantics. The program may have a fixed, objective syntax. But the act of giving meaning to that syntax — whether by talking about the program or by running it — requires something outside the program itself. Any strategy for doing so ultimately depends on social processes.

Consider Python. The Python Reference Manual says that `**` "yields its left argument raised to the power of its right argument."[27] This is an informal specification: it describes the semantics of Python programs using natural language. There are also formal (if unofficial) semantics for Python, which use mathematical notation to define the behavior of Python programs.[28] Or one could run CPython, the most commonly used Python implementation,[29] and confirm that it evaluates `2**3` to `8`.

But wait! Even seemingly innocuous phrases like "raised to the power of" can conceal difficulties. What is 0 raised to the power of 0? Is it 1 because $x^0 = 1$ for all $x \neq 0$? Is it 0 because $0^y = 0$ for all $y \neq 0$? Is it meaningless in the same way that $\text{apple}^{\text{banana}}$ is? This is the kind of question on which mathematicians can disagree.[30] Replacing the

---

[27] GUIDO VAN ROSSUM, THE PYTHON LANGUAGE REFERENCE, RELEASE 3.2.3 49 (Fred L. Drake, Jr. ed., 2012).

[28] *See, e.g.*, Joe Gibbs Politz et al., *Python: The Full Monty*, *in* PROC. OF THE 2013 ACM SIGPLAN INT L CONFERENCE ON OBJECT-ORIENTED PROGRAMMING SYSTEMS, LANGUAGES, AND APPLICATIONS 217 (ACM Press 2013).

[29] *See* ALTERNATIVE PYTHON IMPLEMENTATIONS, https://www.python.org/download/alternatives (last visited May 1, 2019) (calling CPython "the 'traditional' implementation of Python.").

[30] Donald E. Knuth, *Two Notes on Notation*, 99 AM. MATH. MONTHLY 403, 407–08 (1992).

English phrase "raised to the power of" with the mathematical notation "x^y" — as one might in a formal semantics — does not conclusively settle the question, because it is the underlying mathematical concept, not the notation, that is the subject of disagreement. Even CPython is of two minds on the matter. The integer expression `0**0` evaluates to `1`, but the equivalent decimal floating-point expression produces an "Invalid operation" error.[31] This isn't just a Python issue, either. The most recent C standard says that `pow(0.0, 0.0)` is undefined, but many implementations return `1.0`.[32] Is the standard correct? Or is it wrong in the way that an out-of-date dictionary is — no longer reflective of actual usage?

One might reasonably dismiss 0^0 as an unusual, even pathological, example. But it demonstrates in miniature the dependence of technical questions on social ones. Informal specifications, formal semantics, and reference implementations all define the meaning of a program created by humans in terms of *something else also created by humans*. So the meaning of any specific program rests on a foundation of some prior agreement about how to interpret some larger class of programs. Specifications, formal semantics, and reference implementations are not authoritative as a matter of first principles; they are authoritative because people agreed that they are. Why doesn't `2**3` in Python evaluate to `9`? Not because that's what `2**3` inherently means — any more than the seven-letter sequence C-H-I-C-K-E-N inherently means any *gallus gallus domesticus*. In 1991, Guido van Rossum selected `**` as an exponentiation operator for Python and defined its behavior. He could have used `^` instead and made `**` a multiplication operator. If he had, then `2**3` would evaluate to `6`.

But, one might ask, isn't it a logical necessity that $2^3=8$? As long as the Python specification defines `x**y` as x^y, don't the laws of mathematics require that it evaluate to `8` in any correct implementation of Python? There is something to this point, which serves as the foundation of the field of program verification: rigorous standards of proof and truth can be applied to mathematical models of programs. Given a formal semantics of a programming language and a precise specification of a program's operating requirements, it is sometimes possible to produce a logically valid proof that the actual program

---

[31] *See* Devin Jeanpierre, *Issue 23201: Decimal(0)**0 is an error, 0**0 is 1, but Decimal(0) == 0*, PYTHON BUG TRACKER (Jan. 9, 2015, 3:13 AM), https://bugs.python.org/issue23201.

[32] For example, the Apple LLVM 10.0.0 compiler displays this behavior (last tested February 19, 2019 on a MacBook Pro running macOS 10.13.6). I am grateful to Russ Cox for this example.

correctly implements the specification.[33] But there is a crucial step missing: no formal proof is possible that the specification itself corresponds to anything in the outside world.[34] Change the language semantics and all you are left with is an incorrect program and an invalid "proof" of its correctness.

Here is another way of appreciating the point. Consider the Python expression `3/2`. What will happen if you evaluate it? It depends. If you run it in Python version 2.7.15, where / is an integer division operator, it will return `1`. But if you run it in Python version 3.7.1, where / is an exact division operator (and // is the integer division operator), it will return `1.5`. "Python" is not one thing. What we mean when we say "Python" is socially determined.[35] Under some circumstances, we mean Python 2.7.15; under others we mean Python 3.7.1.[36] (If we mean Python 2.7.15, then when we say "the value of the Python expression `3/2`" we refer to `1`, but if we mean Python 3.7.1, when we say "the value of the Python expression `3/2`" we refer to `1.5`. The value of the expression is unambiguous relative to a specific programming language, but that is like saying that the meaning of "chicken" is unambiguous relative to an interpretive convention in which it means any *gallus gallus domesticus*. All the important work is done by the claim that *this* program is written in *that* language. Such claims can only be established by reference to a community of programmers and users.

`2**3` in Python is unambiguously `8`, but that is only because Python users have already agreed on what "Python" is. If they agreed differently, "Python" would be different and so might `2**3`. Collective negotiation over the agreed meaning of "Python" is constantly taking place: in particular, it happens every time there is a new version release. Among other changes, Python 3.7 added a new function called `breakpoint`, but it also made `await` a reserved keyword.[37] Programs that call the `breakpoint` function work in Python 3.7 but not in Python 3.6; programs with a variable named `await` work in Python 3.6 but not in Python 3.7. These changes are debated at immense length on Python developer mailing lists,[38] and each time there is a new release, everyone

---

[33] *See generally* MACKENZIE, *supra* note 1 (describing history of controversies over program verification).

[34] Brian Cantwell Smith, *Limits of Correctness in Computers*, *in* PROGRAM VERIFICATION: FUNDAMENTAL ISSUES IN COMPUTER SCIENCE 275 (Timothy R. Colburn et al. eds., 1993).

[35] So, for that matter, is what we mean when we say "Python 2.7.15."

[36] In linguistic terms, the phrase "Python" is underspecified and requires pragmatic enrichment.

[37] Python Software Foundation, *What's New In Python 3.7* (Elvis Pranskevichus ed.), https://docs.python.org/3.7/whatsnew/3.7.html.

[38] *See, e.g.*, PYTHON-DEV, https://mail.python.org/mailman/listinfo/python-dev.

who is responsible for writing or running Python code decides whether or not to upgrade their version to the latest one. These choices collectively establish the meanings of Python programs — and change those meanings over time. Technical facts depend on socially determined ones.

More precisely, we perceive as fixed technical facts the successful result of coming to social consensus on programming language semantics. A community of programmers and users agrees on a process to extract technical meaning from program text. Developers implement that process on different computers, with different tools, in different contexts. Most of the time, running a program on different implementations will yield the same result. When it does not, technical meaning breaks down.

III. AMBIGUITY IN SMART CONTRACTS

Back to blockchains. We might be able to ignore all of this if smart-contract blockchains never experienced breakdown.[39] But in fact, there are difficulties about the meanings of blockchain programs all the time. I will present four examples, in increasingly dire order.

A. *Oracles*

How does a smart contract observe the world? Suppose, for example, that it needs to release funds from escrow when the seller has delivered a car. The car is a real thing in the real world, not a virtual thing defined by the blockchain VM. The smart contract cannot directly observe it.

The standard solution is to rely on an *oracle* to input real-world data in a form usable by a smart contract.[40] The simplest version of an oracle is simply a trusted user, who is asked to commit transactions verifying that a given event did or not take place, and perhaps supplying some details. This is basically a smart-contract version of an arbitrator or third-party certification. The next step up in complexity is to use a trusted data feed. Trusted software on the blockchain consults some online but off-blockchain data source — like a major financial website's stock quotations — and enters it into the blockchain.[41] The most sophisticated form of

---

[39] *See* TERRY WINOGRAD & FERNANDO FLORES, UNDERSTANDING COMPUTERS AND COGNITION: A NEW FOUNDATION FOR DESIGN (1987) (applying Heiddeger's concept of breakdown to the skew between computer models of the world and the world itself).

[40] *See* ANTONOPOULOS & WOOD, *supra* note 21, at 253–66.

[41] For a sophisticated authenticated data feed solution, see Fan Zhang et al., *Town Crier: An Authenticated Data Feed for Smart Contracts* 270 (2016).

oracle is a consensus oracle: a group of users serve as oracles and the software extracts whatever value they have agreed on. Even simple majority voting can make it harder to corrupt enough involved users to trick a given smart contract, and some consensus oracles use their own consensus protocols, in which the users are rewarded for their participation and for reaching agreement.

But is the oracle correct? We might describe this as a problem of corruption: an oracle that says the car was delivered when it was not is mistaken or lying in the way that a bad judge will be mistaken or lying about a legal contract. Consensus oracles, following the standard blockchain mantra of decentralization, seek to limit corruption by using protocols that encourage correct agreement among the parties. But of course the oracle software has no unmediated access to the truth in the world. Instead, the best its protocols can do is encourage parties to agree — in the hopes that truth will be a more salient focal point than a lie, and that long-term incentives will lead parties to select honest oracles.

The problem of observing the world is also a problem of ambiguity. The world is complex, and contract terms map ambiguously onto the world. An oracle is a way of resolving the ambiguity in how a contract term applies to the infinite variety of factual patterns that could happen in the world. An oracle charged with determining whether the seller in *Frigalament* has performed its obligations resolves any ambiguity about the meaning of "chicken." If the oracle says the seller has performed, then what was delivered was "chicken." If the oracle says the seller has not performed as required, then whatever was delivered was not "chicken."[42]

An oracle's consensus protocol, then, is crucial to how it operates. Single-user oracles and trusted data feeds have simple trust models and consensus protocols; consensus oracles have more sophisticated ones. This leads to two points. The obvious one is that an oracle's resistance to corruption is only as good as its consensus mechanism. The subtler one is that an oracle's ability to resolve ambiguity is only as good as its consensus mechanism.

B. *Upgrades*

Blockchains also upgrade. In 2017, Bitcoin upgraded to implement "segregated witness" (also known as "SegWit").[43] Some data in transactions was moved from one portion of the block to another in a way

---

[42] Allen, *supra* note 5.

[43] Timothy B. Lee, *Bitcoin compromise collapses, leaving future growth in doubt*, ARS TECHNICA (Nov. 9, 2017), https://arstechnica.com/tech-policy/2017/11/bitcoin-compromise-collapses-leaving-future-growth-in-doubt/.

that effectively increased the number of transactions that could fit in each block.[44] The blockchain before SegWit and the blockchain after had different semantics.

Actually, I'm hiding the ball by saying that "Bitcoin upgraded." Blockchains don't upgrade themselves; people upgrade blockchains. Bitcoin's users collectively acted to modify Bitcoin's semantics in ways that would invalidate some transactions. A critical mass of miners announced their support for SegWit, and then on the agreed-upon date started enforcing the new rules. Everyone else went along for the ride. It was just like switching from Python 3.6 to Python 3.7, except that with a blockchain the pressure for consensus is much stronger. Today you can easily find users still happily running Python 3.6, but you will not easily find Bitcoin miners ignoring SegWit.

It's consensus all the way down.[45] The "Bitcoin blockchain" exists only because people agree that it does and what it is. Bitcoin's consensus protocols help coordinate that agreement; indeed, they incentivize it. But the protocols themselves cannot establish their own rule of recognition. A user community can always collectively change or ignore them. This is exactly what happens in an upgrade.

C. *Forks*

Upgrades don't always go smoothly. SegWit was intended (by some users at least) as the first of two linked upgrades to increase Bitcoin's capacity. Following the SegWit upgrade, according to a widely reported-on compromise among various Bitcoin developers, Bitcoin was also supposed to increase its block size from 1 megabyte to 8 megabytes, octupling the number of transactions it could process per block.

This . . . didn't happen.[46] Instead, following the SegWit upgrade, some miners announced they were against the block size upgrade, while others announced they were for it. Discussions and negotiations broke down, and Bitcoin forked into *two blockchains*.[47] One of these blockchains, now known as Bitcoin Cash, increased its block size to 8 megabytes (and then increased it again to 32 megabytes, having

---

[44] *See* Jonathan Cross, *Bitcoin Improvement Proposal 141*, GITHUB (March 10, 2018), https://github.com/bitcoin/bips/blob/master/bip-0141.mediawiki.

[45] Jeffrey M. Lipshaw, *The Persistence of "Dumb" Contracts*, 2 STAN. J. BLOCKCHAIN L. & POL'Y 1, 10 (2018) (smart contracts "have value … simply because there is universal consensus they are what they are").

[46] Lee, *supra* note 42.

[47] Benito Arruñada, *Blockchain's Struggle to Deliver Impersonal Exchange*, 19 MINN. J.L. SCI. & TECH. 55, 73–75 (2018).

established the principle that the block size should grow as needed). The other blockchain, now known as Bitcoin, still has roughly 1 megabyte blocks.[48] The blockchains recognize the same history up until the first >1 megabyte block on Bitcoin Cash, after which they diverge.

Bitcoin and Bitcoin Cash now have different semantics. Is a block valid? The question is unanswerable in the abstract. It can only be answered with reference to a particular blockchain and its user community. A 32-megabyte block is valid according to the agreed-upon semantics of the Bitcoin Cash community, but not according to the Bitcoin community. (It should be obvious that which of them ends up with the "Bitcoin" name is a socially determined fact.)

Blockchain forks are consensus failures. Each blockchain by itself achieves local consensus, but there is no global consensus. Blockchain forks also create explicit ambiguity. The choice of blockchain exposes ambiguity not present when looking at each blockchain by itself. These two facts are inextricably linked, because it is consensus that resolves ambiguity on a blockchain.

Literally anything on a blockchain is subject to the latent ambiguity that the blockchain itself could be upgraded out from underneath it.[49] Whether this actually happens is inescapably political. When there is a disagreement within a blockchain community about a particular upgrade, one of three things could happen. If the pro-upgrade faction backs down, the status quo prevails. If the anti-upgrade faction backs down, the upgrade happens. If neither faction backs down, the blockchain forks. (It should be obvious that which faction, if either, backs down, is an empirical and socially determined fact.)

D. *The DAO*

The DAO — the initialism is short for "distributed autonomous organization" — was a kind of democratic online venture-capital fund.[50] A group of investors planned to join together by using a smart contract on the Ethereum blockchain to manage their affairs, rather than by forming a traditional business organization under the laws of a state. One (imperfect) analogy would be to a venture capital fund operated as

---

[48] I say "roughly" because SegWit complicated the formula for computing block size.

[49] *See* Adam J. Kolber, *Not-So-Smart Blockchain Contracts and Artificial Responsibility*, 21 STAN. TECH. L. REV. 198, 223 (2018) ("So if you agreed to follow the code in the broad sense, then you also agreed to the possibility of a hard fork.").

[50] *See generally* Carla L. Reyes et al., *Distributed Governance*, 59 WM. & MARY L. REV. ONLINE 1 (2017); Usha Rodrigues, *Law and the Blockchain*, 104 IOWA L. REV. 679 (2019).

a general partnership with all of the participants voting on each funding decision.[51]

It flamed out spectacularly.[52] A clever but still unidentified Ethereum user discovered a subtle bug in The DAO contract's code and was able to transfer approximately $60 million worth of Ether to a contract that they alone controlled.[53]

The transfers were quickly noticed, leading to a sharp debate among The DAO and Ethereum users over how to respond.[54] In the end, a large majority of Ethereum users upgraded Ethereum to recognize as valid a new special block with a transaction that unwound The DAO and returned all the funds to the original investors. On this blockchain, which is still known as Ethereum, The DAO and The DAO hack effectively never happened. A minority of users refused to recognize the special block because they considered it contrary to the spirit of smart contracts, blockchains, and Ethereum.[55] On this blockchain, which is known as Ethereum Classic, The DAO and The DAO hack did happen. The two blockchains have different semantics. Indeed, they are incompatible. Transactions now can be entered either on the Ethereum blockchain and conform to its views of which transactions have happened (including The DAO, the hack, and the rollback) or on the Ethereum Classic blockchain and conform to its views (including The DAO and the hack but not the rollback).[56]

---

[51] *See* Christoph Jentzsch, *Decentralized Autonomous Organization to Automate Governance,* https://archive.org/stream/DecentralizedAutonomousOrganizations/WhitePaper_djvu.txt (explaining the implementation of the DAO).

[52] Matt Levine, *Blockchain Company's Smart Contracts Were Dumb*, BLOOMBERG.COM (Jun. 17, 2016), https://www.bloomberg.com/opinion/articles/2016-06-17/blockchain-company-s-smart-contracts-were-dumb.

[53] Nathaniel Popper, *A Hacking of More Than $50 Million Dashes Hopes in the World of Virtual Currency*, N.Y. TIMES (Jun. 17, 2016), https://www.nytimes.com/2016/06/18/business/dealbook/hacker-may-have-removed-more-than-50-million-from-experimental-cybercurrency-project.html.. For technical details, see Phil Daian, *Analysis of the DAO exploit*, Hacking Distributed (Jun. 18, 2016), http://hackingdistributed.com/2016/06/18/analysis-of-the-dao-exploit/.

[54] Joon Ian Wong & Ian Kar, *Everything you need to know about the Ethereum "hard fork,"* QUARTZ (Jul. 18, 2016), https://qz.com/730004/everything-you-need-to-know-about-the-ethereum-hard-fork/.

[55] *The Ethereum Classic Declaration of Independence*, ETHEREUM CLASSIC, https://ethereumclassic.github.io/assets/ETC_Declaration_of_Independence.pdf.

[56] Aaron van Wirdum, *Ethereum Classic Community Navigates a Distinct Path to the Future*, BITCOIN MAGAZINE (Aug. 19, 2016), https://bitcoinmagazine.com/articles/ethereum-classic-community-navigates-a-distinct-path-to-the-future-1471620464/.

The DAO was also (purportedly) governed by a legal contract, although its main job was to defer as much as possible to the smart contract. It stated:

> The terms of The DAO Creation are set forth in the smart contract code existing on the Ethereum blockchain at 0xbb9bc244d798123fde783fcc1c72d3bb8c189413. Nothing in this explanation of terms or in any other document or communication may modify or add any additional obligations or guarantees beyond those set forth in The DAO's code.[57]

In hindsight, this passage is underspecified. The phrase "the Ethereum blockchain" does not uniquely refer. Does it mean Ethereum or Ethereum Classic?[58] It uniquely referred when the contract was drafted, but no longer. It became underspecified — just like any reference to a blockchain could, at any time.[59]

CONCLUSION

We began with three motivations for smart contracts: ambiguity, corruption, and enforcement. It is obvious that protocol changes, forks, 51% attacks, and other consensus breakdowns are a kind of corruption threat to smart contracts. They subject smart contracts to abrogation or alteration at the whims of other blockchain users.[60] It is also obvious that the difficulty of getting people to use a blockchain at all is an enforcement threat. It doesn't matter what a smart contract controlling asset-title tokens on a blockchain says if no one in the physical world pays any attention to the blockchain.

We should also understand the problem of consensus as an ambiguity threat. Natural languages are embedded in communities of people who use and understand those languages. This introduces ambiguity and uncertainty, because people may use and understand the same words in

---

[57] *The DAO - Explanation of Terms and Disclaimer*, THE DAO COMMUNITY (Aug. 3, 2016), https://web.archive.org/web/20160803111447/https://daohub.org/explainer.html.

[58] It should be obvious that the social fact that one of the blockchains is commonly called "Ethereum" and the other is not is relevant but not conclusive in resolving this ambiguity.

[59] *See* Kolber, *supra* note 48, at 222 ("saying that the code is the contract is ambiguous as to precisely what is meant by the code. ").

[60] As I write this, Ethereum Classic was subjected to a $500,000 double-spending attack based on a well-executed deep fork by users who temporarily dominated its mining power. Dan Goodin, *Almost $500,000 in Ethereum Classic coin stolen by forking its blockchain*, ARS TECHNICA (Jan. 8, 2019), https://arstechnica.com/information-technology/2019/01/almost-500000-in-ethereum-coin-stolen-by-forking-its-blockchain/.

different ways. But it also provides a backstop on how badly natural-language contracts can fail. In many cases, the meaning of a contract is clear to a large fraction of people in the relevant linguistic community. If a contract isn't worth the paper it's printed on, it is because of corruption or enforcement problems, not because of ambiguity.

Programming languages appear to reduce linguistic ambiguity. In many cases, they do. Relative to a given implementation, a computer program's meaning is far more definite than a typical natural-language term's meaning. The very process of reducing a term to a formal-language expression requires a degree of precision from its drafters that can itself force them to understand and express their intentions more clearly.

But because programming languages are formal, constructed systems, when the bottom drops out, it can really drop out. The relevant community can redefine the programming language in a way that radically alters the meaning of programs written in it. Smart contracts on a blockchain are particularly vulnerable to this. The same consensus mechanism that keeps them in a local equilibrium can lock them quickly into a new and very different equilibrium — indeed, there are often powerful incentives for users to push the blockchain into a different equilibrium. Blockchain-based smart-contract programming languages don't have continual linguistic drift; they have occasional earthquakes.

In a legal system, the way to change the consequences of contracts is to *change the law*. The natural-language terms in legal contracts still mean what they used to, but their legal effects are different. But on a blockchain, the way to change the consequences of contracts is to *change the semantics*. The programming-language terms in smart contracts mean something different than they used to, and they have different technical effects, and these two differences are the same thing. Interpretation and construction collapse.[61]

This is neither the first nor the last word on ambiguity in smart contracts. I have argued the narrow point that perfect unambiguity is impossible even in theory, because the technical layer ultimately rests on a social one.[62] There is a complementary and broader critique of smart

---

[61] Lawrence B. Solum, *The Interpretation-Construction Distinction*, 27 CONST. COMMENT. 95 (2010). Note that in a *legal* contract incorporating a formal-language term there is still room for construction. As I have argued, these terms are not ambiguous relative to a given formal language; they are ambiguous when there are multiple plausible formal languages in which they could be interpreted and the court (or another legal actor) must select among them. The court might also decide that a term's meaning is clear but nonetheless disregard it for any of the reasons it might disregard a natural-language term. *See, e.g.*, Levine, *supra* note 52.

[62] This is hardly unique to smart contracts or to blockchains. It is a general characteristic of social software. *See* James Grimmelmann, *Anarchy, Status Updates, and Utopia*, 35 PACE L. REV. 135 (2014).

contracts — spelled out best in papers by Karen Levy,[63] Jeremy Sklaroff,[64] and Kevin Werbach and Nicholas Cornell[65] —that even where they do provide unambiguous incorruptible automatic enforcement, this may not be what contracting parties want or need. Writing code is hard, and debugging it is even harder: one advantage of vague and ambiguous natural language is that it is cheaper and faster to negotiate and write down. And sometimes flexibility is good. As Levy explains of legal contracts,

> As such, it can be both operationally and socially beneficial to leave some terms underspecified; vagueness preserves operational flexibility for parties to deal with newly arising circumstances after an agreement is made, and sets the stage for social stability in an ongoing relationship.[66]

And this is to say nothing of the use of smart contracts for socially harmful purposes,[67] the environmental costs of blockchain mining, or the recent blockchain investment bubble.[68]

However, all is not lost for the smart-contract project. Smart contracts cannot be perfectly unambiguous, but they do not need to be perfect to be useful. Socially determined facts are empirically contingent; they are always open to contestation and change. Legal contracts also depend on socially determined facts, and this has not stopped them from having an extremely successful multi-thousand-year run. Much of the time, legal contracts work adequately, despite the ambiguities of natural language. If smart contracts can perform as well or better in even a single domain, they will have a worthwhile role to play.

For better and for worse, blockchains make consensus explicit. The mechanism that holds a blockchain together is the process for agreeing on the next block. Whatever that process yields — in all of its technical and social complexity — is the next block. Every smart contract is therefore only as resilient as its underlying blockchain. Contract law depends on social institutions, particularly those that establish and limit the

---

[63] Karen E.C. Levy, *Book-Smart, Not Street-Smart: Blockchain-Based Smart Contracts and The Social Workings of Law*, 3 ENGAGING SCIENCE, TECHNOLOGY, AND SOCIETY 1 (2017)

[64] Jeremy M. Sklaroff, *Smart Contracts and the Cost of Inflexibility*, 166 U. PA. L. REV. 263 (2017).

[65] Kevin Werbach & Nicolas Cornell, *Contracts Ex Machina*, 67 DUKE L.J. 70 (2017).

[66] Levy, *supra* note 63, at 8. An interesting line of research involves trying to write more deliberately flexible smart contracts. *See, e.g.*, Bill Marino & Ari Juels, *Setting Standards for Altering and Undoing Smart Contracts*, presented at RuleML 2016.

[67] Ari Juels et al., *The Ring of Gyges: Investigating the Future of Criminal Smart Contracts*, *in* PROC. ACM CONF. COMPUTER AND COMMUNICATIONS SECURITY 283 (2016).

[68] *See, e.g.*, Shaanan Cohney et al., *Coin-Operated Capitalism*, 119 COLUM. L. REV 591 (2019) (identifying lack of investor protections in numerous smart contracts).

governments which enforce contracts. Smart contracts depend on social institutions too, particularly those that establish and limit blockchain communities. A blockchain whose governance fails will collapse, fork, or be vulnerable to hijacking. All of these threaten the smart contracts that run on it. There is no escape from politics, because blockchains are made out of people.[69]

---

[69] Curtis Yarvin, *The DAO as a Lesson in Decentralized Governance*, URBIT.ORG (Jun. 24, 2016), https://urbit.org/posts/essays/the-dao-as-a-lesson-in-decentralized-governance/; Steve Randy Waldman, *A Parliament Without a Parliamentarian*, INTERFLUIDITY (Jun. 19, 2016), https://www.interfluidity.com/v2/6581.html; Grimmelmann, *supra* note 62.

ARTICLE

WHY DO PEOPLE AVOID INFORMATION ABOUT PRIVACY?

DAN SVIRSKY†

Why do people keep their head in the sand when making data sharing decisions? There is a widespread intuition, supported by copious research, that people are inconsistent in their behavior around internet privacy. Anger about privacy scandals dominates newspaper headlines, but most people don't change their default privacy settings, even when it's easy to do so. New evidence confirms that this inconsistency is real, and that information avoidance helps drive the inconsistency. This raises a new question: how does information avoidance work? This paper presents a new experimental design to start unpacking how information avoidance operates. There are two main results. First, the experiment replicates existing information avoidance experiments: people who value privacy are willing to deal away their data for small money amounts if given a chance to avoid seeing the privacy consequences of their actions. Second, the experiment shows that while people are comfortable avoiding information about privacy in a passive way, they are not comfortable actively hiding it. These results show that people's ability to keep their head in the sand is fragile: it is a preference people are not willing to exercise conspicuously.

---

† Dan Svirsky, Uber Technologies, Inc. (dsvirsky@uber.com).

INTRODUCTION

Why do people keep their head in the sand when making data sharing decisions?

There is a widespread intuition, supported by copious research, that people are inconsistent in their behavior around internet privacy.[1] Anger about privacy scandals dominates newspaper headlines, but most people don't change their weak, default privacy settings, even when it's easy to do so.[2]

New evidence confirms that this inconsistency is real, and that information avoidance helps drive the inconsistency.[3] Using an

---

[1] *See* Alessandro Acquisti et al., *Privacy and Human Behavior in the Age of Information*, 347 SCI. 509, 510 (2015) (explaining the widespread discrepancies between online privacy attitudes and behaviors); Susan Athey et al., *The Digital Privacy Paradox: Small Money, Small Costs, Small Talk* 17-18 (Nat'l Bureau of Econ. Research, Working Paper No. 23488, 2017) ("Consumers say they care about privacy, but at multiple points in the process end up making choices that are inconsistent with their stated preferences.").

[2] *See, e.g.,* Kevin Lewis et al., *The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network*, 14 J. COMPUTER-MEDIATED COMM. 79, 95 (2008) (finding that only one third of college students using Facebook changed their default privacy settings); Ralph Gross & Alessandro Acquisti, *Information Revelation and Privacy in Online Social Networks*, 2005 ACM WORKSHOP ON PRIVACY IN THE ELEC. SOC'Y 71, 78 (2005) ("We can conclude that only a vanishingly small number of users change the (permissive) default privacy preferences.").

[3] *See* Dan Svirsky, *Why Are Privacy Preferences Inconsistent?* 24 (JOHN M. OLIN CTR. FOR LAW, ECON., & BUS. FELLOWS' DISCUSSION PAPER SERIES, HARV. LAW SCH., Discussion Paper No. 81, 2018) ("This paper presents an experiment that adds to the literature documenting inconsistencies in people's privacy preferences.").

experimental design adopted from research on altruism,[4] this research finds that people are willing to give up nearly an hour's worth of wages to keep their Facebook data private.[5] At the same time, participants in a treatment group are *also* willing to trade their data for 52 cents if given a chance to avoid seeing the privacy implications of their choice.[6] Hence, information avoidance behavior can recreate, in a controlled experimental setting, the pattern of behavior commonly seen in field settings where people are inconsistent about privacy.

This raises a new question: why does information avoidance with respect to privacy online happen?

While the experiment gives strong evidence that people avoid (nearly) costless information about privacy, there are multiple ways to understand this behavior. One possibility is that thinking about losing privacy is inherently unpleasant. There are many topics outside of internet privacy that are inherently upsetting to consider, like cockroaches, death, and one's own moral failings.[7] Avoidance thus reduces the time to consider those unpleasant facets of life. For example, many people eat at restaurants without looking at public health inspection reports on vermin in kitchens. Privacy might be like that.

Another possibility is that even when people know that they *should* care about privacy, they don't really care.[8] Evidence from altruism experiments, for example, demonstrate that people will give money to a Salvation Army volunteer ringing a bell at a supermarket entrance, but

---

[4] *See* Jason Dana et al., *Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness*, 33 ECON. THEORY 67, 70-74 (2007) (describing experimental design of a modified dictator game to test wealth allocation); Zachary Grossman & Joel J. van der Weele, *Self-Image and Willful Ignorance in Social Decisions*, 15 J. EUR. ECON. ASS'N 173, 197-206 (2017) ("[analyzing] a Bayesian signaling model of an agent who cares about self-image and has the opportunity to learn the social benefits of a personally costly action"); Lauren Feiler, *Testing Models of Information Avoidance with Binary Choice Dictator Games*, 45 J. ECON. PSYCHOL. 253, 256-260 (2014) (extending the moral wiggle room experimental design by manipulating the probabilities of different money payoffs). This paper extends the moral wiggle room experimental design in the privacy space in a similar way to the Grossman & van der Weele paper, which also tests the effects of differing the default amount of information presented, albeit in the social preferences space.

[5] *See* Svirsky, *supra* note 3, at 14 ("[F]or these participants, sharing their Facebook profile entails a privacy cost equal to roughly one hour of labor.").

[6] *See id*. (finding that nearly a third of participants chose to share their Facebook profile for 50 cents).

[7] *See* Russell Golman et al., *Information Avoidance*, 55 J. ECON. LIT. 96, 106-07 (2017) (explaining the use of information avoidance as a defense against disappointment).

[8] *Cf.* Christine Exley, *Excusing Selfishness in Charitable Giving: The Role of Risk*, 83 REV. ECON. STUDIES 587 (2016) (demonstrating how participants use risk as an excuse to avoid donating money); Dana et al., *supra* note 4 (showing how people exploit wiggle room to avoid behaving altruistically).

people will also avoid that entrance if there are multiple ways to enter the store.[9] Perhaps in both this domain and privacy, people simply want to *seem* like the type of person who values an important social good (altruism, data security, privacy).[10]

Another alternative is that making a tradeoff between money and privacy is difficult, and people are happy to avoid undergoing this psychic cost.[11] If someone is offered a cup of coffee for $0.25, or for $5.00, she can tell that the first price is somewhat low and the second price somewhat high. The same might not be true for sharing data.

Yet another alternative is that all these explanations hold, to different degrees and with different interaction effects, depending on the person and the context. Perhaps someone wants to *seem* like she cares about privacy, doesn't like thinking about it, and has no real idea what a fair price for data is. All these mechanisms can push her to avoid information. For one person, the first mechanism might dominate. For the same person, the third mechanism might dominate for certain types of data.

This paper presents a new experimental design to start exploring these questions in two steps. First, it replicates the initial two-group experiment on information avoidance in privacy, and second, it adds a third group which has an active choice about hiding information. In the design, participants make decisions about the privacy settings and potential money bonuses for a survey they must complete. They can either complete the survey anonymously or after sharing their public Facebook profile with the survey-taker. Different money bonuses can attach to different privacy settings.

---

[9] *See* James Andreoni et al., *Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving*, 125 J. POL. ECON. 625, 628 (2017) ("When avoidance was easy because only one door had a solicitor, nearly one-third of those intending to pass through the occupied door instead chose to use an unoccupied entrance."). *See also* Edward Lazear et al., *Sorting in Experiments with Application to Social Preferences*, 4 AM. ECON. J: APPLIED ECON. 136, 136 (2012) ("[A]llowing subjects to avoid environments in which sharing is possible significantly reduces sharing."); Stefano DellaVigna et al., *Testing for Altruism and Social Pressure in Charitable Giving*, 127 Q.J. ECON. 1, 1 (2012) (finding that individuals who knew when fundraiser solicitations would arrive at their homes were more likely to avoid the giving scenario).

[10] *See* Christine Exley & Judd Kessler, *Motivated Errors* 1-2 (Harv. Bus. Sch., Working Paper, No. 18-017, 2017) (finding that individuals motivated to act in their own self-interest display behavioral biases yet act more rational when these self-serving motivations are removed).

[11] *See* Cass Sunstein, *Choosing Not to Choose*, 64 DUKE L.J. 1, 1 (2014) ("In part because of limitations of [cognitive resources,] and in part because of awareness of their own lack of information and potential biases, people sometimes want other people to choose for them.").

There are three experimental groups: a *direct* tradeoff group, a *veiled* tradeoff group, and a *choice* tradeoff group. Importantly, there is no real difference in the choices the three groups make. In all three cases, participants decide whether to share their Facebook data for 52 cents, a decision participants can be made aware of if presented the option to view privacy settings. For the *direct* tradeoff group, privacy settings on data sharing are hidden by default; for the *veiled* tradeoff group, privacy settings are visible by default but can be actively hidden; and for the *choice* tradeoff group, privacy settings are visible by default but can be hidden or randomized.

There are three main results. First, the findings replicate the original information avoidance experiment: the *direct* tradeoff group chose privacy 70% of the time, while the *veiled* tradeoff group chose privacy 40% of the time. Second, the findings for the *choice* tradeoff group are directly in between the *direct* and *veiled* groups: participants choose privacy 56% of the time. Third, I find that participants in the *choice* tradeoff group very rarely made the active choice to hide information: they clicked the button to hide privacy settings only 9% of the time, whereas the *veiled* tradeoff group accepted the default of keeping privacy settings hidden 44% of the time.

Taken together, these results shed light on how information avoidance works. People are comfortable avoiding information about privacy, but they are not comfortable actively hiding it. Strangely, even the option of actively hiding makes people less likely to choose privacy.

Section I of the paper discusses current privacy law in the United States as well as the literature on privacy inconsistency and what causes it. Section II unpacks different mechanisms that can drive information avoidance. Section III details the experimental design. Section IV presents the results of the experiment. Section V concludes.

I. PEOPLE'S PRIVACY PREFERENCES ARE INCONSISTENT, AND INFORMATION AVOIDANCE CAN EXPLAIN THIS INCONSISTENCY

A. *Privacy Preferences are Inconsistent*

Extensive experiments and surveys document that people say they value privacy but also give up their data for small amounts of money or convenience.[12] For example, people claim to care greatly about protecting

---

[12] Athey et al., *supra* note 1, at 2 ("Whereas people say they care about privacy, they are willing to relinquish private data quite easily when incentivized to do so."); Leslie John et al., *Strangers on a Plane: Context-Dependent Willingness to Divulge Sensitive Information*, 37 J. CONS. RES. 858, 858 (2011) ("[D]isclosure of private information is responsive to environmental

their data, yet are much less likely to choose a privacy-preserving option if it is listed second on a menu instead of first.[13] Even privacy disclosures that are strikingly clear and scary have limited impact on how much data people give away.[14]

These empirical findings have legal importance because privacy law in the United States relies on a Notice and Choice framework.[15] Firms in the United States can legally harvest data from consumers so long as consumers receive proper notice and agree to the exchange. This approach was first outlined in a 1973 report by the U.S. Department of Health, Education and Welfare.[16] The reliance on notice and voluntary consent was a departure from how privacy law originally developed. Before the rise in internet commerce and telecommunications, privacy concerns in transactions between non-state actors were governed by tort law.[17] As internet transactions have come to dominate private data, contract law principles have come to increasingly govern privacy law.[18] Since privacy is governed by consumer choice, the well-documented fickleness in how consumers make privacy decisions has policy importance.

There are exceptions to the Notice and Choice framework. Banks send annual privacy notices because of the Gramm-Leach-Bliley Act.[19] Doctors require patients to sign an extra form because of the Health Insurance Portability and Accountability Act.[20] Websites ask users if they are older

---

cues that bear little connection . . . to objective [privacy] hazards."). *See, e.g.,* Alessandro Acquisti et al., *What is Privacy Worth*, 42 J.L. STUD. 249, 268-69 (2013) (discussing how individuals make inconsistent decisions in privacy contexts in part because of default privacy settings). *Cf.* Adam Chilton & Omri Ben-Shahar, *Simplification of Privacy Disclosures: An Experimental Test* 566 (Coase-Sandor Working Paper Series in Law and Economics, No. 737, 2016) (describing the failure of simplified privacy disclosures to effect meaningful change in participants' behavior in disclosing private information).

[13] Athey et al., *supra* note 1, at 12 ("[W]hen wallets that would maximize privacy from the public are not listed first, students are 13% less likely to select them . . . .").

[14] Chilton & Ben-Shahar, *supra* note 12, at 541 ("[B]est-practice simplification techniques have little to no effect on respondents' comprehension of the disclosure, willingness to share personal information, and expectations about their rights.").

[15] *See generally* Chilton & Ben-Shahar, *supra* note 12, at 572-73 (discussing the emphasis in American privacy law on giving proper notice to consumers).

[16] Records Computers and the Rights of Citizens, U.S. DEP'T OF HEALTH, EDUCATION AND WELFARE, SUMMARY AND RECOMMENDATIONS, xxx-xxxii (1973).

[17] *See, e.g.*, Richard A. Posner, *The Right of Privacy*, 12 GA. L. REV. 393, 410 (1978) (discussing the general features of tort-based commercial privacy law); William L. Prosser, *Privacy*, 48 CAL. L. REV. 383, 389 (1960) (highlighting the four types of privacy torts). *Cf.* Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890) (articulating the need for common law to grow to cover an individual's right 'to be let alone' and provide a remedy for invasions of privacy by the press").

[18] There is more stringent regulation of certain consumers and certain industries.

[19] 15 U.S.C. §§ 6801(b), 6805(b)(2) (2000).

[20] 42 U.S.C. § 1320d-2(d)(2) (2000).

than 13 -- not 18, not 12, not 16 -- because of the Childrens Online Privacy Protection Act.[21] Outside the United States, there is even more stringent regulation. The European Union has started enforcing the General Data Protection Regulation, which imposes stronger consent requirements for data collection, forces firms to delete personal data at a consumers request, and allows for fines up to 4% of a firms global revenue. [22] Hence, more muscular regulation does exist, and the political will for it is increasing. But in the United States, such regulation is the exception.

The standard explanations for the inconsistency in measures of how people value privacy are bounded rationality and revealed preference.[23]

Under bounded rationality, people are unaware of how much data they are emitting or they struggle to value privacy. The latter may be because privacy is abstract, or because privacy costs are inchoate and uncertain, both in scope and timing.[24] Either way, people do not fully understand what is at stake. As a result, when deciding whether to exchange privacy for something more easily quantifiable, like money or convenience, small frictions may play an outsized role in decision-making.[25] This line of scholarship draws on classic findings from psychology and economics, like the endowment effect and framing effects, to explain peoples fickle privacy preferences.[26]

Under the revealed preference explanation, people give up privacy simply because this maximizes their utility.[27] People trade privacy for money, or convenience, because this is what they actually prefer, regardless of what they say. If information has some cost, then consumers decision to avoid privacy information is itself an illustration of revealed preference.

---

[21] 15 U.S.C. § 6502(b)(1)(D) (2000).

[22] Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Advancement of Such Data, and repealing Directive 95/46/EC, 2016 O.J. L 119/1, Art. 83 § 5.

[23] *See* Svirsky, *supra* note 3, at 2 (noting that scholars point to bounded rationality or cognitive bias to explain inconsistency in privacy choices).

[24] *See* Acquisti, *supra* note 12, at 251-52 (stating that privacy violation costs are amorphous and difficult to assess even when quantifiable).

[25] *See id.,* at 267 (explaining that data from one experimental design shows subjects were five times more likely to choose privacy when the trade-off to not doing so was framed as an opportunity to add to an initially gifted amount of money as opposed to retaining the entirety of the initially gifted amount).

[26] *See id.,* at 252 (showing empirically that endowment effects and order preferences affect privacy valuations).

[27] *See* Athey, *supra* note 1, at 4 (“The second policy our results document is that there is a disconnect between stated privacy preferences and revealed preference, but that revealed preference is actually closest to the normative preference.”).

For either explanation – revealed preference or ignorance – more information is better. If its costless, people will always opt for better information about privacy settings. More recent research suggests a third explanation.

B. *Information Avoidance Can Explain this Inconsistency*

Recent research demonstrates that information avoidance can explain inconsistency in people's privacy decisions.[28]

There is a robust literature from psychology and economics on information avoidance.[29] While economists typically model information as an intermediate good[30] – i.e., valuable only because it helps us achieve ends – scholars in psychology and economics increasingly recognize that people sometimes behave as if information has emotional valence.[31] This leads to a recognition that more information is not always better.

This pattern of information-avoiding behavior is important across information-sharing domains. People will give money to a non-profit when a fundraiser goes door to door; many of the same people will find an excuse not to answer the door if they are warned ahead of time that a fundraiser is coming.[32] In the health sector, one study found that 27% of intravenous drug users at risk of HIV who got tested for the disease did not return to the clinic to see their results,[33] even though knowing ones HIV positive status can lengthen ones life. People avoid information that upsets them, even if in theory it should help them make a more optimal decision.

Indeed, such behavior appears to be at play in privacy choices as well. Svirsky (2018) demonstrates that even people who value keeping data private at willingness-to-pay ("WTP") prices of several dollars are willing to give up their data at nominal prices if they can avoid immediately seeing

---

[28] *See* Svirsky, *supra* note 3, at 24 (concluding that information avoidance may drive privacy decisions).

[29] *See e.g.,* Golman, *supra* note 7, at 110 (summarizing the literature in economics and psychology related to regret aversion and optimism maintenance).

[30] *See generally* Posner, *supra* note 17 (analyzing the economics of information from an individual perspective to improve privacy analysis); George Stigler, *The Economics of Information*, 69 J. POL. ECON. 213 (1961) (modeling the ascertainment of market price in order to improve economic organization techniques).

[31] *See* Emily Oster et al., *Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease*, 103 AM. ECON. R. 804, 806 (2013) (analyzing the impact of an individual's expectations in determining whether to undergo genetic testing).

[32] *See* DellaVigna et al., *supra* note 9, at 3 (finding that individuals who knew the exact time at which fundraising solicitors would arrive at their homes were more likely to not open the door to the solicitors).

[33] Patrick Sullivan et al., *Failure to Return for HIV Test Results Among Persons at High Risk for HIV Infection*, 35 J. ACQUIRED IMMUNE DEFICIENCY SYNDROMES 511, 515 (2004).

the result of their decision.[34] In the experiment, participants completed a survey after first deciding whether to do the survey anonymously ("high privacy") or after giving their public Facebook profile data to the survey-taker ("low privacy") for a bonus.[35] A control group chose between {0 cents, high privacy} and {50 cents, low privacy}.[36] Roughly two thirds of participants opted for "high privacy", and in follow-up treatments, most participants refused to opt for "low privacy" until offered at least $2.50.[37]

In a treatment group, participants faced a choice between {0 cents, privacy option A} and {50 cents, privacy option B}.[38] They knew that the two privacy options were randomized so that "privacy option A" could be "high" or "low" privacy with a 50% chance, and vice versa.[39] Importantly, participants in the treatment group could click to reveal the privacy options *before choosing*, at no monetary cost.[40] If they click a button, they know that they will then either see {0 cents, high privacy} and {50 cents, low privacy} as in the control group, or they will see a more obvious choice between {0 cents, low privacy} and {50 cents, high privacy}.[41]

The key finding was that hiding the potential privacy settings behind a veil – even when removing the veil is costless – causes a drop in people's willingness to keep their data private.[42] The percentage of people who refused 50 cents to stay anonymous dropped from 67% in the control group to 40% in the treatment group.[43]

Importantly, this treatment effect that occurs for decisions between a money bonus or privacy does not replicate for decisions between two privacy settings, both associated with money bonuses (with the second money bonus drawn from the distribution of people's willingness-to-pay prices for privacy). When a second money bonus is hidden by a costless veil, participants do not evince the same willingness to engage in information avoidance.

While the treatment effect is large and robust – it was documented across four experimental rounds across several months in a sample size of over 1000 subjects[44] – it leaves open important questions of what specific mechanism drives information avoidance behavior.

---

[34] Svirsky, *supra* note 3, at 24.

[35] *Id.*, at 6.

[36] *Id.*, at 9.

[37] *Id.*, at 13.

[38] *Id.*, at 9.

[39] *Id.*

[40] *Id.*

[41] *Id.*

[42] *Id.*, at 21.

[43] *Id.*, at 15.

[44] *Id.*, at 11.

II. MULTIPLE MECHANISMS CAN EXPLAIN INFORMATION AVOIDANCE BEHAVIOR

People engage in information avoidance when making privacy decisions.[45] That is, they avoid looking at low-cost information about how their data will be shared, even when they value keeping their data private. But why?

This section begins by modeling how a participant makes choices in the information avoidance experiment before then discussing competing mechanisms to explain the treatment effect and how the model can be extended to incorporate these mechanisms.

A. *Modeling the Wiggle Room Decision*

An agent is making a tradeoff between a payoff and an uncertain cost. For example, she might be deciding whether to download the Uber app, knowing that her data might be sold or her location tracked. Suppose the app brings some value v and a privacy cost c which occurs with probability π. Then, in a standard expected utility model, her utility is as follows:

$$u(\pi) = v - \pi c$$

Now suppose that there is a psychic cost to potentially losing privacy. The thought of something upsetting is itself upsetting. Let the function $\psi()$ map π onto disutility, with

1. $\psi(\pi) > 0 \; \forall \, \pi \in [0,1]$
2. $\psi'(\pi) > 0 \; \forall \, \pi \in [0,1]$

The first condition says that the possibility of something upsetting is in itself upsetting. The second condition says that the agent gets more upset as the upsetting possibility becomes more likely -- a 100% chance of getting an electric shock upsets the agent more than a 10% chance. Throughout, I will assume that $\psi(0) = 0$. If $\psi(\pi) = 0 \; \forall \, \pi \in [0,1]$, we are in the case of classical preferences, where information is only valuable for instrumental reasons but has no valence in and of itself.

The person's utility function now incorporates psychic costs:

$$u(\pi) = v - \pi c - \psi(\pi)$$

---

[45] *Id.*, at 24.

In cases where a persons information is fixed – she has a belief about π but can do nothing to change this belief – psychic costs are akin to increasing the cost of a harm, albeit in a potentially non-linear way. The comparative statics are straightforward: more psychic costs means an individual is more likely to avoid an action. Where psychic costs will generate more interesting departures from standard models is in decisions over how much information to collect before making a decision.

How does a participant make choices in the experimental design described above? Consider what happens when a participant gets what is commonly described as "wiggle room" – or the chance to make a choice where they give up their data without directly seeing that they are giving up their data. That is, they can choose a higher monetary payoff while still telling themselves that they might be keeping their data private.

In the control group, the participant makes a direct tradeoff between money v and the privacy cost c and the psychic cost of losing privacy with near-certainty, $\psi(\pi_H)$, where π_H is close to 1. She chooses to keep her privacy if the monetary payoff v is lower than the cost (psychic or otherwise) of losing privacy:

$$u(\pi) = v - \pi_H c - \psi(\pi_H) < 0$$

In the *veiled* tradeoff treatment, the participant first has to make a choice about whether to lift a veil, or whether to remain ignorant and take a higher payoff.

Suppose the privacy options are randomized, so that if she lifts the veil, then with probability p she will discover she is in the baseline condition (more money means less privacy), and with probability $1 - p$ she will discover she is in an easy situation where she can get more money *and* keep her privacy. If she remains ignorant, her payoff is:

$$v - pc - \psi(p)$$

In words, she gets the value v with certainty, but undergoes a privacy cost c with probability p and has a psychic cost $\psi(p)$. Consider a participant who would have opted for privacy in the control treatment, so the value of privacy is higher than the monetary payoff v. If she lifts the veil, then her expected payoff is:

$$p(0) + (1 - p)v = (1 - p)v$$

That is, with probability p she will face a tradeoff between privacy and money and will keep her privacy as before, yielding payoff 0; the rest of the time she will get a payoff without any privacy costs (psychic or instrumental).

In a classical preferences world – one where people value information solely because it helps them make better choices, and where information has no attendant psychic costs – $\psi(\pi) = 0 \; \forall \, \pi \in [0,1]$, and any agent who chose privacy over money in the baseline treatment will always choose to lift the veil. Why? If, in baseline, she chose privacy over money, that means

$$\begin{gathered} v - \pi_H c - \psi(\pi_H) < 0 \\ v - \pi_H c < 0 \\ v < \pi_H c \\ v < \pi_H c < 1 \cdot c \\ v < c \end{gathered}$$

In the *veiled* tradeoff treatment, she lifts the veil if

$$\begin{gathered} (1 - p)v + p(0) > v - pc - \psi(p) \\ v - pv > v - pc \\ -pv > -pc \\ v < c \end{gathered}$$

The last line is true by assumption. Putting the conclusion into plain language: if clicking to reveal the privacy settings is costless, then anyone who values privacy more than the money bonus would make it their business to *see* which privacy options they were agreeing to. The monetary bonus is simply not worth the risk of giving up data.

In sum, the model demonstrates that the experimental result in Svirsky (2018) cannot be obtained from classic preferences, so long as the cost of clicking to reveal is minimal. The following subsections turn to different mechanisms that *can* explain the wiggle room result.

B. *Mechanism: Thinking about Privacy Losses is Upsetting*

One explanation for the experimental results is from a model of anxiety in which people are upset by probabilistic harms.[46] Importantly, the magnitude of the psychic harm need not be a linear function of the probability of the harm. Unlike in classic expected utility theory, when a 100% chance of something good is exactly twice as nice as a 50% chance of the same reward, psychic costs can have different shapes.

If someone has convex psychic costs – e.g., a 1%, or 2%, or 50% chance of harm are all treated like a 0% chance – then the wiggle room result can be obtained.

Again, assume the agent chooses privacy over money in the baseline treatment. That means:

$$v - \pi_H c - \psi(\pi_H) < 0$$

In the information avoidance treatment, she remains ignorant if:

$$\begin{gathered} (1-p)v + p(0) > v - pc - \psi(p) \\ -pv > -pc - \psi(p) \\ pc - pv + \psi(p) < 0 \\ p(c - v) + \psi(p) < 0 \\ p(v - c) - \psi(p) > 0 \end{gathered}$$

Unlike before, an agent with psychic costs might opt for privacy in the *direct* tradeoff treatment, but choose to remain ignorant in the *veiled* tradeoff treatment.

For this to happen, we need two conditions: $v > c$ and a functional form for $\psi(\cdot)$ which is convex. This means that in the baseline case, psychic costs are what is driving the agent to opt for privacy. At the same time, her psychic costs are relatively low when losing privacy is uncertain: a 0.01% chance of losing privacy, or a 1%, or 10%, or 50% chance – all these feel distant from a 100% chance. Consider the classic Star Wars quote when an anxious C-3PO warns the heroic Han Solo about the odds of successfully navigating an asteroid field.[47] Solo shouts back: "never tell me the odds."[48] Here, Solo is like an agent with convex preferences over probabilistic harms. Whether

---

[46] *See* Botond Koszegi, *Health Anxiety and Patient Behavior*, 22 J. HEALTH ECON. 1073, 1074 (2003) (describing a model in which a patient's utility function is defined by her expectations about her future physical outcomes).

[47] STAR WARS EPISODE V: THE EMPIRE STRIKES BACK (Lucasfilm Ltd. 1980).

[48] *Id.*

the probability of crashing is 0.01 or 0.99, he needs to ignore the danger and treat all probabilities as if they are zero. He wants to remain ignorant.

C. *Mechanism: Signaling*

Another explanation for the experimental result is signaling: people care about privacy, but they also care about *seeming* like they care about privacy.[49]

This drives a wedge between the *direct* tradeoff group and the *veiled* tradeoff group, because members of the *veiled* tradeoff group can take the monetary bonus without explicitly sacrificing privacy. In the *direct* tradeoff group, taking the monetary payoff and rejecting privacy comes with a signaling cost of showing (either to herself or an observer) that the participant does not value privacy. In the *veiled* tradeoff group, meanwhile, taking the monetary payoff without looking at the privacy choices carries no such signaling cost.

In the model, this would mean that the psychic cost of losing privacy depends on how observable her choice is. The monetary value v is the same across groups, but in the *veiled* tradeoff, if the cost of knowingly losing privacy is c, then the cost of losing privacy without being aware of doing so is $c_0 < c$. Hence, some people who would choose to keep their privacy in the *direct* tradeoff treatment ($c > v$) would take the money and not click to reveal the privacy settings in the *veiled* tradeoff treatment ($c > v > c_0$). If she gives up privacy without a (costless) veil, the psychic cost is imposed. If there is a veil, then the psychic cost is lower.

D. *Mechanism: Psychic Choosing Costs*

Some scholars posit that the act of making a choice imposes costs.[50] The *direct* tradeoff group faces a direct choice between money and privacy, which may be difficult if privacy costs are inchoate or hard to measure. The *veiled* tradeoff group, meanwhile, can opt out of a difficult choice by refusing to consider it. The veil, then, creates a treatment effect by letting people avoid choosing costs.

In the model, this works like signaling in reverse. People in the *direct* tradeoff group face a psychic cost of losing privacy (due to the difficulty of making the choice). If they give away their privacy, they lose cost c, but the

---

[49] Zachary Grossman & Joel J. van der Weele, *Self-Image and Willful Ignorance in Social Decisions,* 15 J. EUR. ECON. ASS'N 173, 176 (2017) (concluding that endogenous signaling is one driver of behavior in social situations).

[50] Cass Sunstein, *Choosing Not to Choose*, 64 DUKE L.J. 1, 40 (2014) (noting that active choice imposes a large burden on the chooser, unlike passive acceptance of a default).

act of choosing imposes a psychic calculation cost c_{choose}. People in the *veiled* tradeoff group, meanwhile, face no such cost unless they click to reveal the privacy settings. The result is that there exist participants who opt to remain anonymous in a control group setting but refuse to unveil – and then give up their privacy – in a treatment group setting. That is, $c > v$, so they choose privacy in the *direct* tradeoff treatment, but c_{choose} is large enough that it is not worth clicking to reveal the privacy settings and choosing to remain anonymous.

E. *Mechanism: All of the Above*

None of the explanations above are mutually exclusive. They may also be operative, to different degrees in different people. They may interact, so that cases with high choosing costs are also ones where signaling is more powerful. The interactions themselves may differ across people. Hence, the treatment effect might occur for one participant because she finds it unsavory to think about privacy losses, she has a hard time choosing, *and* she really only cares about *seeming* like she cares about privacy. For another participant, the treatment effect might hold because she actually does care about privacy but hates thinking about it, so she does not click to reveal in the *veiled* tradeoff treatment. A different participant might actually want to seem like she is *not* worried about privacy, but also has a hard time making tradeoffs between privacy and money, so the mechanisms would work in opposite directions.

In short, while the existence of a treatment effect from the wiggle room experimental design has been demonstrated for privacy decisions, many mechanisms might be at play. The remainder of this paper turns to exploring these mechanisms with an additional experimental treatment.

III. EXPERIMENTAL DESIGN AND EMPIRICAL APPROACH

304 participants were recruited on Amazon Mechanical Turk to take a short survey about health and financial status. All participants were informed that before doing the survey, they would make decisions about the size of a bonus payment, to be received upon completion, and the privacy settings of the survey.[51] The experiment was conducted on January 7, 2019. The sample of participants was limited to those in the United States.

Research increasingly suggests that, for the purpose of social science experiments, Mechanical Turk users are a reliable sample. One might be

---

[51] The median wage in the study was \$15.33 (based on a median payment of \$1.02 for a median completion of 3:59 seconds).

concerned about how findings in this population translate to others. Because the setting is Mechanical Turk, one can assume that the sample is quite computer literate and also is comfortable completing short (mundane) tasks for a low wage. However, research suggests that these external validity issues are not of first-order importance. Irvine (2018) replicates three experiments using in-person labs, national online platforms, and Mechanical Turk, and finds that the results are constant across samples.[52] Nonetheless, as with any experiment, the sample of participants is important to keep in mind when interpreting results.

After recruitment, the timeline of the experiment consists of three stages: instructions, privacy settings, and a survey.[53] First, participants were shown an initial introductory screen giving an overview of their participation. Participants were told that they would take a survey, but while everyone would take the same exact survey, each participant would be given a choice between two privacy options. They could opt for high privacy, in which case their survey answers would be anonymous. Or, they could opt instead for low privacy, in which case they would click a "Log in with Facebook" button at the top of the survey. This meant that the survey-taker would see, in addition to the participants survey answers, her public Facebook profile (including profile picture, name, and gender) and her email address. Participants who chose low privacy would not be allowed to finish the survey until they logged in.

After the instructions stage, participants chose their privacy settings. After completing the privacy settings stage, participants completed the survey.

The privacy measure in the experiment – whether to share Facebook information – has three advantages: it is a real decision, it is a realistic one, and it is an important one. First, participants who give up their privacy in this experiment must actually give over their profile data, so the choice is not a hypothetical one. Nor is it a behavior that can be faked; unlike other privacy experiments, which measure privacy as a persons willingness to answer an intrusive question, a participant in this experiment cannot pretend to give up privacy without actually giving anything up.[54] Second, the decision is a realistic one. The "Log in with Facebook" button is a

---

[52] *See* Krin Irvine et al., *Law and Psychology Grows Up, Goes Online, and Replicates*, 15 J. EMP. LEG. STUD. 320, 343-44 (2018) (demonstrating the key difference that Mechanical Turk users were significantly more attentive than other samples).

[53] For detailed study instructions, please email the author at dsvirsky@uber.com.

[54] Even if participants have a fake account they can use -- Facebook works hard to limit such behavior, but is not always successful -- handing over a fake account involves some cost. Doing so means the experimenter can link a fake Facebook account to a Mechanical Turk account (and the answers in the survey), which makes the fake account less effective.

ubiquitous part of the internet - many websites allow people to log in with their Facebook (or Google) account rather than with the website itself. Hence, it is a choice people routinely make: should I engage in online activity in a way that is linked to my Facebook profile or not? Third, the decision has important public policy implications, as suggested by the Cambridge Analytica scandal.[55]

Each person was randomized into one of three treatments during the privacy settings stage: the *direct* tradeoff treatment, the *veiled* tradeoff treatment, and the *choice* tradeoff treatment. The exact format of the privacy choice made in each of the treatments can be seen in Figure 1 (*direct* tradeoff), Figure 2 (*veiled* tradeoff), and Figure 3 (*choice* tradeoff).[56]

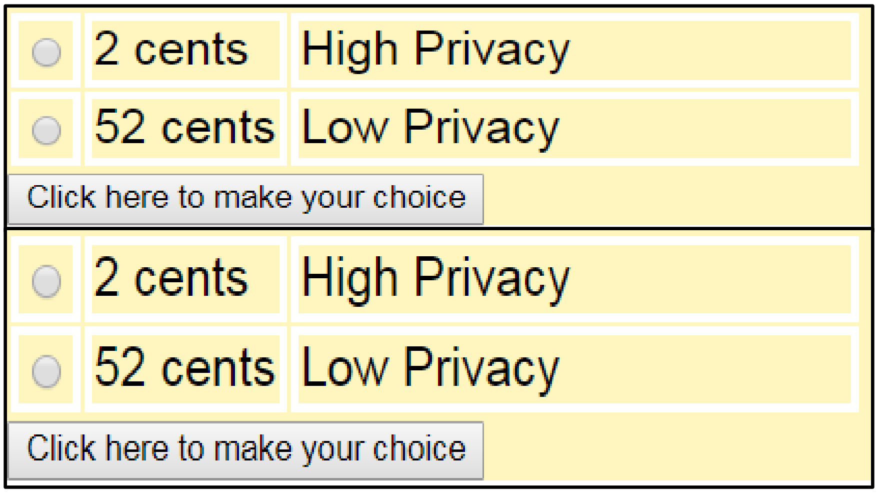


Figure 1: In the *direct* tradeoff group, participants are aware that they are choosing between privacy and money, as both settings are visible by default.

---

[55] The privacy measure is less ecologically valid in the sense that it is about sharing data with a researcher, rather than a corporation or a government. It could be that people are more comfortable sharing data with an academic researcher than with Facebook or a police department. The opposite could also be true. In any case, this would cause all three experimental groups to change how they value privacy, but not impact them differentially. One interesting note is that in the Cambridge Analytica scandal, the malicious actors who harvested data posed as academic researchers.

[56] One contribution of this paper is replication. The first two groups – the *direct* and *veiled* tradeoff groups – face a decision identical to that in Svirsky, *supra* note 3. That paper finds a treatment effect of information avoidance. This paper expands on that paper by adding a third treatment group and is also an opportunity to replicate and retest the initial findings, which is vital for the health of scholarly disciplines that rely on sound experimental findings. *See*, *e.g.*, Irvine et al., *supra* note 52.

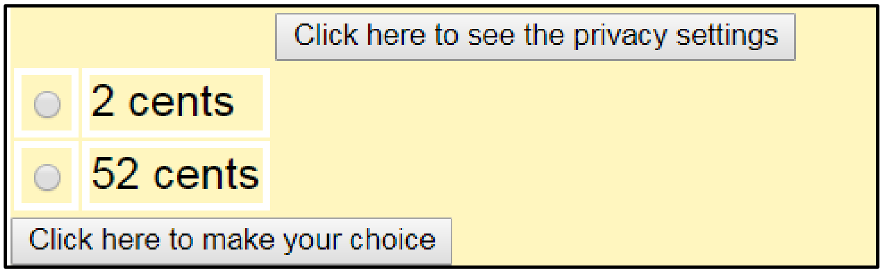

Click here to see the privacy settings
○ 2 cents
○ 52 cents
Click here to make your choice

Figure 2: In the *veiled* tradeoff group, participants choose between privacy and money. The privacy setting is hidden by default but can be revealed instantly and costlessly.

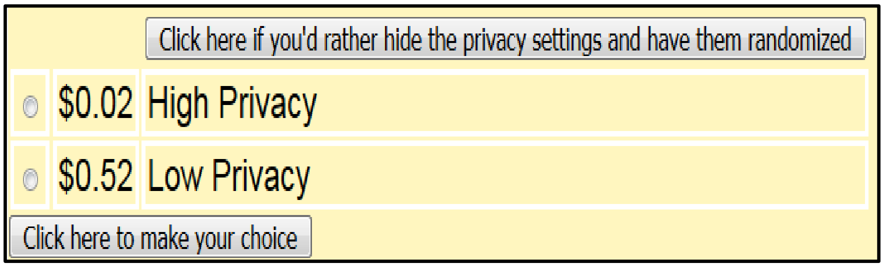

Click here if you'd rather hide the privacy settings and have them randomized
○ $0.02 High Privacy
○ $0.52 Low Privacy
Click here to make your choice

Figure 3: In the *choice* tradeoff group, participants choose between privacy and money. The privacy setting is visible by default but can be hidden (and randomized) instantly and costlessly.

In the *direct* tradeoff treatment, participants only made one decision: a direct choice between a 2-cent bonus and privacy option A or a 52-cent bonus and privacy option B. The privacy options were randomized so that half the time, participants faced a degenerate choice between { more money, more privacy } and { less money, less privacy }. The other half of the time, participants faced a true tradeoff between money and privacy.

In the *veiled* tradeoff treatment, participants faced the same decision as in the *direct* tradeoff treatment, but the privacy setting was initially hidden. Participants had to click to reveal the column describing the privacy settings, and there was a 50% chance that the higher money bonus would mean losing their anonymity.[57]

In the *choice* tradeoff treatment, participants faced the same layout as in the *direct* tradeoff treatment, but they had the option of clicking a button to hide (and randomize) the privacy settings. Upon clicking, the privacy

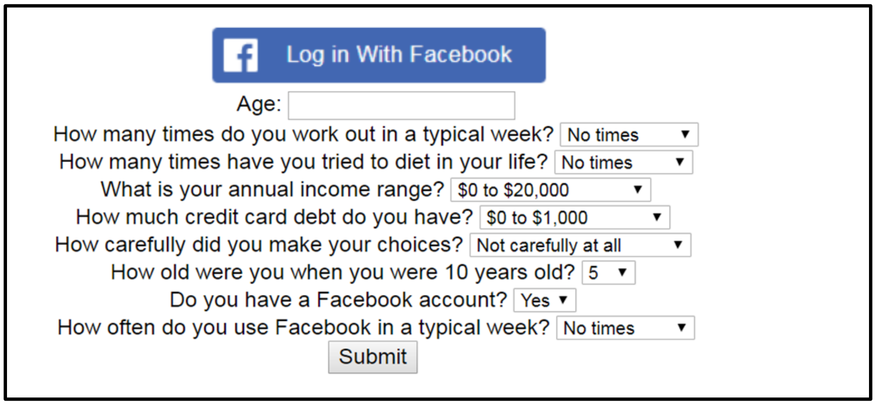
---

[57] Note that for both groups, there was a 50% chance of facing a degenerate choice between { more money, more privacy } and { less money, less privacy }. These decisions cannot tell us about how much a person values privacy, so they are omitted from the main analyses below.

settings were hidden, and participants faced the same layout (and choice set) as the participants in the *veiled* tradeoff treatment.

In sum, all participants faced the same choice, but depending on random assignment, they faced a different default layout. Some saw everything – money and privacy options – and had to choose directly. Some started off by seeing everything but through active choice could hide the privacy settings. Some started off with privacy settings hidden, and through active choice, could have revealed these settings. Hence, if clicking is costless, there should be no difference between the three groups.

After completing the privacy stage, all participants completed a nine-question survey, shown in Figure 4. Five questions covered demographics, health, and financial topics. These questions asked about the persons age, the number of times they exercise in a week, the number of times they have attempted to diet in their life, their annual income, and their credit card debt. The survey also included two questions to check comprehension. One asked "How old were you when you were 10 years old?" with a dropdown menu with several options, including 10. Another directly asked "How carefully did you make your choices?" with three options: "Not carefully at all", "I thought about it a little", and "I was very careful". Two questions asked whether participants had a Facebook profile and how often they used Facebook. After submitting the survey, participants were finished.



Age:
How many times do you work out in a typical week? No times ▾
How many times have you tried to diet in your life? No times ▾
What is your annual income range? $0 to $20,000 ▾
How much credit card debt do you have? $0 to $1,000 ▾
How carefully did you make your choices? Not carefully at all ▾
How old were you when you were 10 years old? 5 ▾
Do you have a Facebook account? Yes ▾
How often do you use Facebook in a typical week? No times ▾

Figure 4: Screenshot of the survey that each participant completed. The "Log in With Facebook" button only appears if the participant opted to share her Facebook info. If she instead opted for anonymity, the button would not be included.

The demographic questions were selected somewhat arbitrarily, since they were not the focus of the experiment. The goal was to find questions that were somewhat intrusive (that implicate some privacy concerns) without being offensive. The comprehension and Facebook questions help to interpret any results. If a participant does not have a Facebook account, it is hard to interpret her privacy choices. Similarly, if the treatment effect is driven by people who fail comprehension questions, or who use Facebook rarely, then this is informative in understanding what drove any treatment effect.

The user interface for the experiment was coded using HTML and Javascript, which ensured that the "reveal button" would work instantaneously -- without a page refresh. When a user clicked the reveal button, Javascript code changed the visibility setting of the hidden column from hidden to visible. The hidden column would therefore become visible immediately. The users choices and data were sent to a MySQL database using PHP code.[58]

Even though the choice is essentially 50 cents vs privacy, it is more accurate to note that this is a 50-cent *bonus*. The participants are foregoing 50 cents, not actually giving away any of their pre-experimental wealth. There is extensive behavioral economics literature noting the distinction between losses and gains.[59] This point is broadly important but has little relevance here. Participants would be more likely to opt for money over privacy if the monetary change were a loss rather than a gain, but this would affect all three experimental groups equally.

IV. RESULTS

The results are organized as follows: Section A gives summary statistics and balance checks, while Section B shows the primary findings – the average treatment effects as compared to how often each group chose to remain anonymous, as well as how often the *veiled* and *choice* tradeoff groups clicked to hide or reveal the privacy settings.

---

[58] All code is available on request from the author and includes survey instructions, experimental module coding, and the raw data. Contact the author for the ZIP file: dsvirsky@hbs.edu.

[59] *See, e.g.*, Botond Koszegi & Matthew Rabin, *A Model of Reference-Dependent Preferences*, 121 Q. J. OF ECON 1133, 1134 (2006) (expounding on prospect theory as applied to consumer behavior).

A. *Summary Statistics*

There were no systematic demographic differences between the treatment groups, as expected given the random assignment. Of note, 90% of participants reported having a Facebook account, and the median participant used Facebook three times per week.

	Direct Tradeoff (N = 108)	*Veiled* Tradeoff (N = 109)	*Choice* Tradeoff (N = 87)	P-Value
Age (years)	34.15 (10.24)	33.41 (9.112)	34.82 (10.40)	0.61
Diet Attempts in Lifetime (0 – 4)	2.102 (1.646)	1.954 (1.512)	2.517 (1.547)	0.04
Exercise Workouts in a Typical Week (0 – 4)	2.213 (1.454)	2.358 (1.385)	2.276 (1.476)	0.76
Annual Income (0 – 4)	1.519 (1.196)	1.385 (1.053)	1.494 (1.160)	0.66
Credit Card Debt (0 – 4)	0.815 (1.051)	1.018 (1.097)	0.839 (1.066)	0.32
Has Facebook (0,1)	0.917 (0.278)	0.890 (0.314)	0.874 (0.334)	0.61
Weekly Facebook Use (0 – 4)	2.491 (1.568)	2.541 (1.549)	2.793 (1.526)	0.36

Table 1: Summary statistics. Standard deviation reported in parenthesis. With the exception of age, which is reported in years, each variable is categorical. Hence, an answer of 1 for credit card debt corresponds to a range of $1000 to $2000 in debt.

B. *Main Results: Average Treatment Effects*

Do either of the two treatments lead people to choose privacy more often? When given the option to hide or reveal information, do participants do so?

In all three treatments, participants faced the same choice: participants were offered 52 cents to share their Facebook data, or 2 cents to preserve their anonymity. The only difference was in the default information presented. Nonetheless, I find a significant impact on people's willingness to sell their data. Roughly 70% of people in the *direct* tradeoff group opted to remain anonymous. In the *veiled* tradeoff group, only 40% remained anonymous. These numbers are almost identical to those found in Svirsky (2018).[60] Meanwhile, however, participants in the *choice* tradeoff group – who saw both money and privacy settings but had the choice to hide the privacy settings – were halfway between the other groups. Roughly 56% of participants in the *choice* tradeoff group opted to remain anonymous: less than in the *direct* tradeoff group, but more than in the *veiled* tradeoff group. Figure 5 presents the results graphically. Table 1 shows the results of regression models where {ended up staying private} is the binary dependent variable, and there are indicator variables for the *veiled* and *choice* tradeoff groups. Each column presents a different sample, each one representing a robustness check.

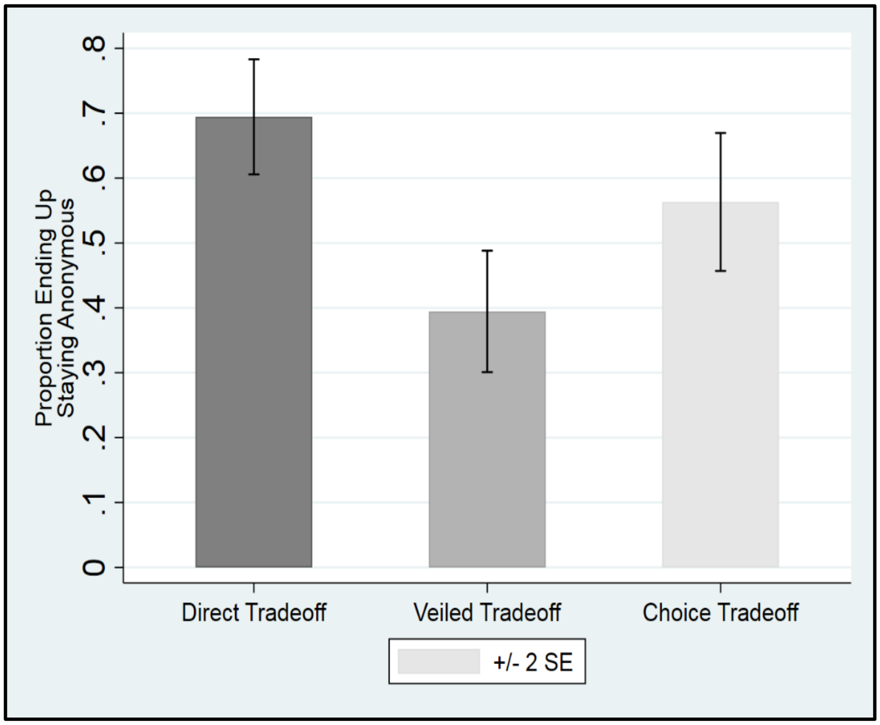


Figure 5: This figure shows the proportion of participants who ended up remaining anonymous for 2 cents instead of sharing their Facebook profile for 52 cents, for the *direct* tradeoff group (N = 108), the *veiled*

---

[60] *See* Svirsky, *supra* note 3, at 1 (finding that online survey participants had to make the same choice whether to share their Facebook profile data with the survey taker in exchange for a higher payoff).

tradeoff group (N = 109), and the *choice* tradeoff group (N = 87). These results exclude all participants who, by randomization, faced a degenerate tradeoff of 52 cents and high privacy vs 2 cents and low privacy. Therefore, for the *veiled* tradeoff group, anyone who chose the higher money option is counted as having chosen 50 cents over anonymity, regardless of whether they clicked to reveal the privacy setting before making their decision.

	Entire Sample	Excludes People Who Fail Comprehension Check	Excludes People Who Did Not Answer Carefully	Excludes People w/o a Facebook account
Veiled Tradeoff	-0.30*** (0.06)	-0.29*** (0.07)	-0.29*** (0.07)	-0.32*** (0.07)
Choice Tradeoff	-0.13* (0.07)	-0.13* (0.07)	-0.13* (0.07)	-0.13* (0.07)
Constant	0.69*** (0.05)	0.71*** (0.05)	0.70*** (0.05)	0.69*** (0.05)
N	304	264	294	272
Adjusted R^2	0.06	0.05	0.06	0.07

Table 2: Regression models of average treatment effect. The dependent variable is a binary variable for whether the person ended up remaining anonymous. There are indicator variables for the *veiled* and *choice* tradeoff groups, so the constant represents the proportion of participants in the *direct* tradeoff group who ended up remaining anonymous. Each column uses a different subset of the sample in order to provide robustness checks. *** $p < 0.001$, * < 0.10.

Participants in the *choice* tradeoff group by and large did *not* hide information about privacy. In the *choice* tradeoff group, only 9% of participants made an active choice to hide (and randomize) the privacy settings before making a choice. In the *veiled* tradeoff group, where privacy settings were hidden by default (but could be revealed), 45% of participants made a choice without seeing the privacy information. This difference in proportions is statistically significant (Fisher's Exact $p < 0.001$).

CONCLUSION

This paper explores why people avoid information about privacy when making data sharing decisions. Existing work demonstrates that even when privacy settings are easy to read – even just two words long – people who otherwise would pay several dollars to remain anonymous are happy to avoid looking at the settings and take a 50-cent bonus. This paper solidifies this behavior. It finds that while people are happy to avoid information that is already hidden, they are not likely to actively hide information that is in front of them to begin with. At the same time, the option of hiding information makes people marginally more likely to sell their data, even if they do not choose to hide the privacy settings.

The results give more support to certain mechanisms of information avoidance than others. Theories that rely on signaling are consistent with the data presented here. If people care more about *seeming* like they value privacy, then they might take the money if given plausible deniability (as in the *veiled* tradeoff group), but not go so far as to actively hide information (as in the *choice* tradeoff group), as such a choice would signal a willingness to care little about privacy.

Theories that posit that people simply prefer not to think about privacy, or prefer not to choose, are less consistent with the data. If these mechanisms explain information avoidance, then people would opt to simplify their choice if given the option.

The results also suggest that current U.S. privacy law – centered around giving consumers better information – may be difficult to achieve in practice. There is considerable scholarship and policy experimentation around giving people simpler, more effective disclosures.[61] Simpler disclosures is likely a good thing: if it gives people more information at lower costs, this should improve welfare. At the same time, if many people choose to avoid information about privacy, then better disclosures will not be as effective as a classical economics model would suggest. If societies want people to end up with more privacy, it will be difficult to do so by relying on individuals to seek out the information they need and choose accordingly.

---

[61] *See*, *e.g.*, Chilton & Ben-Shahar, *supra* note 12, at 1 (describing the failure of simplified privacy disclosures to effect meaningful change in participants' behavior in disclosing private information); Corey Ciocchetti, *The Future of Privacy Policies: A Privacy Nutrition Label Filled with Fair Information Practices*, 26 J. MARSHALL J. COMP. & INFO. L. 1 (2008-2009) (discussing standardization of labels to force all e-commerce homepages to conspicuously post their privacy practices).

ARTICLE

JUSTIFYING THE EFFICACY OF CONTRACT DISCRIMINATION

HOSEA H. HARVEY†

In recent years, the insights of behavioral law and economics scholars have improved the efficacy of various forms of contract-regimes through substantive legal reforms ranging from the CARD Act to a revamped RESPA. These insights and reforms attempted to optimize consumer choice architecture and enhance overall consumer decision-making utility, primarily by a combination of new information-deployment techniques and various consumer nudges, in both standardized paper formats and online. But much more can be done to build on these insights and improve decision-making in this space – in order to maximize utility for historically marginalized groups. This Article argues that as more traditional commercial transactions move online, they can be more easily customized to directly engage consumers by directly taking into account a consumer's race and other demographic factors.

Encouraging discrimination in contract formation comes with potential barriers and costs. Certain federal and state regulations prohibit the acquisition and use of such data. Privacy experts caution against the expansive use of online tools and algorithms designed to inferentially gather such data. Consumer demand for racially customized online interactions is uncertain. And, the potential for corporate misuse of such data, to discriminate in harmful ways, is possible. But these concerns should be measured against potential market benefits and can be addressed by rigorous data analysis of completed contracts. In certain regulated consumer markets, digital platforms that would seek to acquire race data and customize contracts would be required to permit regulators to evaluate whether such contract disclosures and contract terms were discriminatory. Ultimately, in the absence of a more transparent and honest dialogue about the present acquisition and use of such information

---

† Associate Professor of Law, Temple University, Beasley School of Law.

in online contracts, an unregulated market can utilize such information at will and without scrutiny – which runs the risk of harming consumers and carries unknown benefits.

INTRODUCTION: COUNTING RACE AND MAKING IT COUNT

As individual consumers, we respond to, utilize, and learn from advertising, marketing, disclosure, and information regimes, in print and online, on a daily basis. As traditional consumer contract markets have moved to digital formats, contract-making has become both more personalized and more automated based upon the engagement of personal preferences.[1] Many of these consumer markets are regulated with a light touch, if at all, and thus the full extent of a typical market-seller's use of a customer's personal data to structure terms is unclear.[2]

---

[1] *See* Rory Van Loo, *Digital Market Perfection*, 117 MICH. L. REV. 815 (2019) (predicting far greater automation of consumer transactions based on personal preferences). *See also* Joshua A.T. Fairfield, *Smart Contracts, Bitcoin Bots, and Consumer Protection*, 71 WASH. & LEE L. REV. ONLINE 35, 38 (2014).

[2] For this reason, it is difficult to assess whether and how the use of such data impacts consumer utility in those markets. To the extent that consumers are harmed in those markets, the interventions described here would prove costly. However, to the extent that the transparency proposed here identifies differential and negative market effects for certain consumer segments, such evidence could serve as the empirical basis to expand regulatory oversight of "light-touch" markets.

But, in other consumer markets, whether online or in-person, the federal government structures the methods by which market actors engage consumers from the earliest stages of the contract formation process, such as with mandated consumer disclosures for prescription drugs or consumer credit products.[3]

This Article argues that corporations subject to these additional oversight regimes should be encouraged to gather socio-demographic information for print and online transactions and customize contracts based upon that information. The decision-enhancing framework underlying consumer disclosure law finds its original source in law and economics principles, namely that individuals, once identified and provided with information, will "rationally optimize their choices, given their preferences, information, and the incentives they face."[4] The Truth in Lending Act ("TILA") was enacted with this basic premise.[5] Moreover, information's rationalizing effect should protect and enhance the interests of consumers by positioning them to make welfare-optimizing decisions. Policymakers are increasingly relying on digital intermediaries to play that rationalizing role through disclosures aimed at machines. If those machines are supposed to help consumers, and if a consumer's interests are tied to their socio-demographic background, why shouldn't corporations be able to incorporate and utilize this information in ways consistent with decision-enhancing principles?[6]

As leading scholars from other areas have recognized, race, gender and other factors can be excluded from evaluating and informing a

---

[3] This principle can be broadly applied to a range of government sanctioned information dissemination regimes. Here, the information of particular value is consumer disclosure, specifically with respect to consumer finance. One of the earliest modern examples of this strategy, of course, is pursuant to the National Traffic and Motor Vehicle Safety Act of 1966, as amended, (49 U.S.C. 30112(a), 30115). National Traffic and Motor Vehicle Safety Act of 1966, 49 U.S.C. §§ 30112(a), 30115 (1966). Under the Act, a motor vehicle manufactured for sale in the United States must have affixed a label certifying compliance with various mandates and applicable standards. The label, among other things, must identify the vehicle's manufacturer, its date of manufacture, the Gross Vehicle Weight Rating or GVWR, the Gross Axle Weight Rating or GAWR of each axle, the vehicle type classification (e.g., passenger car, multipurpose passenger vehicle, truck, bus, motorcycle, trailer, low-speed vehicle), and the vehicle's Vehicle Identification Number or "VIN." 49 C.F.R. § 567.4 (2013).

[4] *See* Ryan Bubb & Richard H. Pildes, *How Behavioral Economics Trims Its Sails and Why*, 127 HARV. L. REV. 1593, 1602 (2014). As explained *infra*, recent efforts by BLE scholars to improve such laws have necessarily challenged this assumption.

[5] Matthew Edwards, (quoting ELIZABETH RENUART & KATHLEEN E. KEEST, TRUTH IN LENDING § 1.1.1, at 33 (4th ed. 1999) (describing TILA as "Congress's effort to guarantee the accurate and meaningful disclosure of the costs of consumer credit and thereby to enable consumers to make informed choices in the credit marketplace").

[6] *See* Rory Van Loo, *Rise of the Digital Regulator*, 66 DUKE L. J. 1267 (2017) ("The administrative state is leveraging algorithms to influence individuals' private decisions.")

decision-making process, but "from a technical perspective . . . this approach is naïve. Blindness to a sensitive attribute has long been recognized as an insufficient approach to making a process fair."[7] The resultant product is "insufficient to assure fairness and compliance with substantive policy choices."[8] Thus, to maximize the effectiveness of consumer transactions in a digital era, we may need to focus less on how such transactions affect consumers generally and more on how such transactions are designed for, utilized by, and impact marginalized consumer groups, particularly racial and ethnic groups.[9]

There are critiques of this approach, discussed later in Section III, including whether such a regime implicates privacy concerns and whether government's encouragement of "discrimination" in this context violates core moral or ethical principles. But it is useful to begin with a third critique about the underlying theory and evidence for such an approach: by and large, we do not know how, when, and why we might expect consumers from different groups to respond differently to particular types of contracts.[10] However, this absence of evidence is partly because much consumer contract and behavioral law and economics ("BLE") disclosure-centered scholarship has often swept socio-demographic variables like race under the behavioral rug, exacerbating this empirical dilemma.

A. *The Importance of Evaluating Racial Differences in Commercial Law Scholarship*

When we believe race matters, as an independent explanatory or causal variable to differentiate consumer interests, experiences, or

---

[7] *See* Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 685 (2016) (discussing ECOA's Reg B. prohibitions – and their failings – within a larger framework about debiasing machine algorithms).

[8] *Id.*

[9] Though not the primary focus of this Article, financial literacy and education regimes similarly suffer from a lack of focus on the information needs of marginalized groups. *See* Lauren E. Willis, *Against Financial Literacy Education*, 94 IOWA L. REV. 197, 228-29 (2008) (arguing that remedies must be context-specific to be impactful). *See also Final Report President's Advisory Council on Financial Capability*, FINAL REPORT, 10 (2013), http://www.treasury.gov/resource-center/financial-education/Documents/PACFC%20Interim%20Report%20-%20January%2018,%202012.pdf (stating that recommendations should "take into account the particular needs of traditionally underserved populations (e.g., women, minorities, low- and moderate-income consumers, and the elderly)").

[10] *See, e.g.*, Dalié Jiménez, D. James Greiner, Lois R. Lupica & Rebecca L. Sandefur, *Improving the Lives of Individuals in Financial Distress Using a Randomized Control Trial: A Research and Clinical Approach*, 20 GEO. J. ON POVERTY L. & POL'Y 449 (2013).

contract outcomes, academics and government policy makers should encourage private actors and government regulators to acquire that information and then to utilize it to inform or improve law and public policy.[11] Similarly, when we see an absence of effort to gather or analyze or deploy such information, it sends a clear message that the underlying social phenomenon or policy problem either should not or does not implicate race or racial justice matters. In short, when race matters, government and private actors should count it, analyze it, and use the resulting knowledge and information to reduce disparities and improve public welfare.[12] Within academia, we expect the same level of effort.[13] While encouraging the acquisition of racial demographics for commercial transactions may not always yield an obvious net utility,[14] there is general agreement that racial difference permeates a variety of consumer contract regimes in a variety of ways.[15] But, in the context of recognizing the role of race in communicating with consumers, legal scholars in other fields are far ahead of commercial law academics.[16]

B. *Generic Consumer Contract Approaches*

We know the fallacy of the central assumption of traditional law and economics approaches to individual decision making — that consumers are rational maximizers of their strategic goals.[17,18] Drawing upon social science research, BLE scholars proved that human behavior and

---

[11] This assumes that government generally seeks to make such policies better, rather than worse.

[12] *See* Ming Hsu Chen & Taeku Lee, *Reimagining Democratic Inclusion: Asian Americans and the Voting Rights Act*, 3 U.C. IRVINE L. REV. 359 (2013) (advocating for broader data gathering and data analysis by race to improve efficacy of voting rights laws).

[13] *See, e.g.,* Gregory S. Parks, *Toward a Critical Race Realism*, 17 CORNELL J. OF L. AND PUB. POL. 683 (2008) (encouraging critical race theorists to deploy social science data analysis methodologies when analyzing law and public policy problems). *See also* Devon W. Carbado & Daria Roithmayr, *Critical Race Theory Meets Social Science*, 10 ANN. REV. OF L. & SOC. SCI. 149 (2014) (explaining how social science research offers critical race theory scholars a useful methodology).

[14] *See* Jonathan D. Kahn, *Patenting Race*, 24 NATURE BIOTECHNOLOGY, Nov. 2006, at 1349 (2006) (raising concerns about utilizing race as a variable when petitioning the government in patent and drug-approval spaces).

[15] *See, e.g.*, Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REG. 547, 579-80 (2016) (observing that sellers' algorithms have the potential to lessen some forms of racial discrimination and exacerbate others).

[16] *See, e.g.,* Dayna Bowen Matthew, *Race, Religion, and Informed Consent — Lessons from Social Science*, 36 J. OF L., MED. & ETHICS 150 (2008) (gathering and analyzing empirical and historical data to re-contextualize the role of race and ethnicity in informed consent agreements).

[17] *See, e.g.,* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2008).

[18] Bubb and Pildes, *supra* note 4.

decision-making consistently differs from that of the rational actor.[19] Accordingly, such scholars contend that policymakers should legislate with an eye toward "minimiz[ing] the individual mistakes that create behavioral market failures and . . . mitigate their negative consequences."[20] With respect to one such area of law, disclosure law regimes, they believe that government-mandated disclosures – provided through market intermediaries, should "focus . . . on helping [real] people help themselves."[21]

"To date, the work in BLE has been surprisingly circumscribed,"[22] and, by assuming that "many" or "most" consumers exhibit the *same* behavioral biases that impact rational decision-making in the *same* way, much BLE literature falls victim to the presumptive errors also made in law and economics theory.[23] In other words, BLE improves upon rational-actor models by anticipating predictable forms of decision-making errors, but also assumes that all consumers act irrationally in *consistent* ways or make imperfect decisions using information in a predictably imperfect manner. By baselining these models, and then subsequent policy and law derived therefrom, on a "universal" person, this suggests, but does not explicitly state, that the consumer is racially white – and male.[24]

Thus, when relying on this assumption of a mythical universal generic consumer, there is less need to engage the efficacy, impact, or value of incorporating consumer demographic differences in consumer contracts, as there is no accompanying theoretical explanation for why such consumers would be expected to process information differently or yield different utility from similar decisions. Therefore, if this core BLE assumption were true, disclosure models or digital "smart contracts" created for a singular class of "irrational" consumers would prove effective in reducing noise, increasing decision-making efficiencies, and leaving consumers better off – as a whole – than without the information.

But what if a contract formation's utility function varies across a range of socio-demographic groups? For example, one of the few large-

---

[19] See, e.g., Thaler and Sunstein, *supra* note 17.

[20] *Bubb and Pildes, supra* note 4, at 1605.

[21] *Id.* at 1604.

[22] *Id.*

[23] Colin Camerer, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue & Matthew Rabin, *Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism,"* 151 U. PA. L. REV. 1211, 1219 (2003). ("[W]e can divide consumers into two types: those who are boundedly rational (in the sense described above) and those who are fully rational; and that (2) a fraction, p, of consumers fall into the boundedly rational category."),

[24] *See generally* IAN HANEY LOPEZ, WHITE BY LAW (1999).

scale credit-granting disclosure experiments confirms that the quality and visibility of consumer disclosures specifically matter for vulnerable high-risk populations, who "are rate sensitive only if the interest rate information is prominently disclosed."[25] Similarly, the empirical relationship between contract-formation choices, credit-card profiles, and certain socio-demographic information (especially race) is still uncertain, though suggestive of group-based differences.[26] Therefore, it is consistent with the limited scholarship that exists that certain population sub-groups could react sub-optimally (or simply differently) to proposed contracts and terms that others (a majority) use efficiently and rationally.[27] If this is true, even rational economic decision making - mediated through information - is not uniformly distributed.[28] Yet, contract disclosure and formation defaults, particularly as they are operationalized – in style, language, and substance – are effectively white. But in the digital world, the provision of information and disclosure could instead be focused on customizing the contract formation process to maximize the economic-utility and welfare of population sub-groups.[29] If the provision of information can be designed to maximize the utility of sub-groups such that it serves to enhance not

---

[25] *See* Bruno Ferman, *Reading the Fine Print: Credit Demand and Information Disclosure in Brazil*, 62 MGMT. SCI. 3534 (2015) (conducting a large-scale credit card disclosure experiment in Brazil and finding, in part, that "most borrowers are highly rate-sensitive, whether or not interest rates are prominently disclosed in marketing materials. An exception is high-risk borrowers, for whom rate disclosure matters.")

[26] A number of studies have shown correlation between (perceived) race and credit scores, suggesting that, in fact, there are clear financial health differences by population, although the causes of such differences remain largely unknown. *See*, *e.g.*, EEOC, *May 16 Hearing Record* (statement of Adam T. Klein) (citing the 2000 Freddie Mac National Consumer Credit Survey) (correlation between race and credit score); Bd. of Governors of the Fed. Reserve Sys., REPORT TO THE CONGRESS ON CREDIT SCORING AND ITS EFFECTS ON THE AVAILABILITY AND AFFORDABILITY OF CREDIT 80–81 (2007), available at http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf. (finding African-Americans and Latinos have lower credit scores than other racial/ethnic groups); Matt Fellowes, Brookings Inst., CREDIT SCORES, REPORTS, AND GETTING AHEAD IN AMERICA 2 (2006), https://www.brookings.edu/research/credit-scores-reports-and-getting-ahead-in-america (showing correlation between percentage of racial minority residents and a U.S. county's average credit score).

[27] *See, e.g.,* David Hoffman, *From Promise to Form: How Contracting Online Changes Consumers*, 91 N.Y.U. L. REV. 1595 (2016) (demonstrating, in part, that socio-demographic differences in consumer groups are associated with differing views about the contract formation process and the implications for contract breach).

[28] *See*, *e.g.*, Mintel, HISPANIC FINANCES AND FINANCIAL SERVICES (2009) (finding higher Latino race-differential response rates to the question "I know nothing about financial services/investments.")

[29] *See* Shmuel I. Becher, Yuval Feldman, and Orly Lobel, *Poor Consumer(s) Law: The Case of High-Cost Credit and Payday Loans* in LEGAL APPLICATIONS OF MARKETING THEORY, Jacob Gersen & Joel Steckel, eds., Cambridge University Press (2019, Forthcoming).

only the general public welfare but the welfare and maximal utility of sub-groups as well, the result would be a more optimal outcome than the status-quo.[30] And while some scholars have innovatively encouraged a "performance test" of various disclosure and contract-formation regimes to further enhance consumer utility, [31] such tests still compare the utility maximizing effect of such differentiated regimes on the outcomes for the consumer population *as a whole* – rather than distinct sub-groups. Instead, the approach here contemplates a variant of what others have described as a consumer finance "randomized control trial," in which digital experimentation with how consumer contract disclosures are provided and the efficacy of particular terms will allow for a real-time gathering of evidence of what works best – holding the socio-demographics of the reference-group constant.[32] Such experimentation can be achieved faster through digital contracts and the use of online platforms, which would also allow for a more rapid aggregation of evidence about the efficacy of this approach.

However, recognizing that this approach is not without its weaknesses, this Article responds to those who may believe this approach to contract formation will do more harm than good.[33]

I. CRITIQUES OF DISCLOSURE LAW AND A CRITIQUE OF THOSE CRITIQUES

Because the purpose of disclosure law, as mentioned, is to enable rational decision-making by consumers, the law and economics movement—and its critics—have comprised the foundation of scholarly commentary on the impact and efficacy of consumer disclosure laws, to the exclusion of those focused on achieving racial justice. Most

---

[30] *See* Kroll, et. al., *supra* note 7, at 682 (acknowledging privacy concerns and reviewing potential discriminatory effects in using algorithms but suggesting that "there may be cases where allowing an algorithm to consider protected class status can actually make outcomes fairer. This may require a doctrinal shift, as, in many cases, consideration of protected status in a decision is presumptively a legal harm.")

[31] Lauren E. Willis, *Performance-Based Consumer Law*, 82 U. CHI. L. REV. 1309, 1316 (2015) (footnotes omitted).

[32] *See*, *e.g.*, Jimenez et. al., *supra* note 10, at 470 (describing a large-scale mixed-methods research study to gauge the effectiveness of financial health interventions via a "consumer incentive to undergo financial counseling, an offer of attorney representation, and the two treatments in combination.")

[33] *See*, *e.g.*, Lea Shepard, *Toward A Stronger Financial History Antidiscrimination Norm*, 53 B.C.L. REV. 1695, 1711-718 (2012) (questioning empirical assumptions associated with an employer's use of job applicants' financial histories and arguing, in part, for a more robust anti-discrimination norm with respect to consumer credit-information regimes due to the potential racially disparate impact associated with their use.)

particularly, the behavioral economics movement has dedicated significant resources "to tak[ing] the core insights and successes of economics and build[ing] upon them by making more realistic assumptions about human behavior . . . [seeking to provide] a better description of the behavior of the agents in society and the economy."[34] Such scholars have drawn upon psychological and sociological scholarship that has not only acknowledged the countless cultural and environmental factors that impact how individuals respond to contract-formation stimuli, but have also embraced them as variables to predict future behavior.

But, this scholarship has failed to acknowledge, explain, or even identify whether – and how – race, ethnicity, and other socio-demographics impact – or are impacted by – the very disclosure regimes such scholars seek to change. Thus, celebrated law scholars in this space whose work is rightly lauded for its general behavioral insights have remained curiously silent about whether sub-group differences exist in responding to optimizing information distribution in similar welfare enhancing ways.[35]

Current scholarship denotes disclosure policy as a political device *designed* to remedy information asymmetries in the market place.[36] Nonetheless, those endorsing such laws and regulations cannot ignore evidence identifying the deficiencies in disclosure law's implementation. Those questioning the merit of the current regime predominantly point to "empirical evidence and theories regarding consumer behavior," "deficiencies of the disclosures themselves," as well as "the [in]ability or [un]likelihood [that] consumers . . . use the

---

[34] Christine Jolls, Cass Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471, 1487 (1998).

[35] Pathbreaking in a variety of ways, such works simply fail to engage the role of consumer race, ethnicity, and culture (as well as sex), as if these variables are not factors in how consumers receive, process, or act on disclosure. *See, e.g.,* Margaret Jane Radin, BOILERPLATE: THE FINE PRINT, VANISHING RIGHTS, AND THE RULE OF LAW (2012); Oren Bar-Gill, SEDUCTION BY CONTRACT (2012); Omri Ben-Shahar and Carl E. Schneider, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2015). Perhaps one reason these scholars and others see disclosure as so ineffective is precisely because of its lack of experimental differentiation with terms across consumer sub-groups. Compare this absence of discussion in commercial law literature with the engagement of race variables and critical theory in other substantive fields, such as health law. *See, e.g.,* Khiara Bridges, Terence Keel & Osagie K. Obasogie, *Introduction: Critical Race Theory & the Health Sciences*, in Symposium Critical Race Theory & the Health Sciences, 43 AM. J. OF LAW & MED. 179 (2017).

[36] Matthew A. Edwards, *Empirical and Behavioral Critiques of Mandatory Disclosure: Socio-Economics and the Quest for Truth in Lending*, 14 CORNELL J.L. & PUB. POL'Y 199 (2005).

information"[37] among the reasons for its ineffectiveness. Perhaps the absence of much socio-demographic information is the real cause.

Scholars have also identified the possibility of supply-side issues – that the complexity and volume of information may render it meaningless to a confused or overwhelmed consumer. First, evidence has shown that complexities in the law itself can stunt the compliance efforts of regulated entities.[38] While this is true for all populations, there is increasing evidence that demonstrates that the complexity is particularly salient for population sub-groups more so than the general population and that it causes members of these sub-groups to make, on average, more inefficient decisions with the same information.[39] Similarly, an oft-cited defect of contract-formation that behavioral theorists recognize is "information overload," the argument that "consumers [are] cognitively unable to cope with the voluminous nature of the mandated . . . disclosures."[40] With respect to TILA, subsequent to its 1980 emendation, scholarship dedicated to dissecting this particular issue somewhat subsided.[41] Nonetheless, "home mortgage borrowers [are still] . . . buried in paper, with little guidance as to which documents contain the most crucial information to facilitate credit decision-making."[42] Most (if not all) consumers find this problem familiar, as they attempt to process overwhelming amounts of information online to make the most efficient contracting choices.

II. IMPLEMENTATION

The following examples build on a premise not universally shared by BLE scholars – that particular population sub-groups may exhibit non-random decision-making errors with respect to evaluating contract disclosures and terms.[43] This non-random error distribution can result

---

[37] *Id.* at 204.

[38] *Id.*

[39] *See, e.g.,* Kleimann Communication Group, KNOW BEFORE YOU OWE: POST-PROPOSAL CONSUMER TESTING FOR THE CFPB OF THE SPANISH AND REFINANCE INTEGRATED TILA-RESPA DISCLOSURES (2015) (discussed infra).

[40] Edwards, *supra* note 36, at 221.

[41] Edwards attributes three explanations to this: (1) "the application of information overload theory to legal regulation has been subjected to a significant amount of scrutiny and criticism," (2) wariness surrounding "advocating a position that might lead towards recommendations of less disclosure for consumers," and finally, (3) that the amended regulations arguably "ameliorated the worst of TILA's overload problems." *Id.* at 222.

[42] *Id.* at 223.

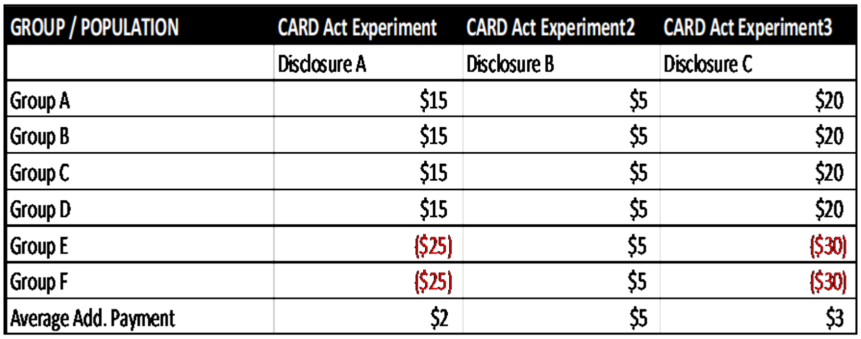
[43] *See*, *e.g.*, Shmuel I. Becher et. al., *supra* note 29 (explaining that certain BLE assumptions about consumer financial behavior are not evenly associated with certain groups – and may be particularly flawed for marginalized groups.)

from a variety of causes. Here, I'll focus on three of them: (a) language barriers, (b) unique socio-demographic differences in the processing or utilization of information, or (c) a non-randomly distributed lack of engagement with information. Each of these root causes, if proven, would lead to a different set of proposed information-based solutions particularly suitable for digital transactions, in part because the costs typically associated with the deployment of socio-demographically varied disclosure and terms in print form would be substantially reduced. We can think about these solutions as falling within three broad frameworks: (a) improving contract-formation utility by clarification, (b) improving contract-formation utility by addition of group-relevant topics, and (c) improving contract-formation utility by individuation. Each of these solutions requires the gathering and use of socio-demographic variables and robust evidence testing. Each of them is also particularly easy to test and execute for digital contracts, because a controlled experiment incorporating the modification of contract terms or disclosure language and evaluating differential responses can be accomplished at higher speed and lower cost than creating, distributing, and evaluating responses from differentiated printed and distributed versions of the same material.

Example: Online Credit Card Applications

In order to prove a claim that the utility of information might vary across subgroups, one would prefer empirical validation from real-world evidence. Lacking that in this case forces speculation – on both sides. On the one hand, BLE scholars assume without proof that no differences exist, and their models reflect this. Here, as a thought experiment, let me illustrate a universe where the utility of contract-formation information *does* vary across subgroups – in order to postulate what we might do in response were this to be so.

First, let us imagine a scenario where, prior to the implementation of the Credit Card Accountability and Disclosure ("CARD") Act's revised disclosure and education model, a simulated online test was run using three versions of that model. One of the goals of the new model embedded within the CARD Act is to educate consumers that an increase in the amount paid per month will reduce the overall cost of a medium-term extension of credit.[44] The goal is to increase monthly payments through the education function of disclosure, which over time will enhance the welfare of consumers because they will spend less money for the extension of credit over time. Let us speculate that three different versions of online disclosure were tested with that goal in mind, across particular demographic sub-groups with the same number of participants, with results as follows:[45]



GROUP / POPULATION	CARD Act Experiment	CARD Act Experiment2	CARD Act Experiment3
	Disclosure A	Disclosure B	Disclosure C
Group A	$15	$5	$20
Group B	$15	$5	$20
Group C	$15	$5	$20
Group D	$15	$5	$20
Group E	($25)	$5	($30)
Group F	($25)	$5	($30)
Average Add. Payment	$2	$5	$3

In the above experimental framework, the primary goal is to maximize the additional monthly payment of the consumer population as a whole in order to reduce the long-term cost of credit. In that scenario, Disclosure B is the optimal choice, because it maximizes the average additional payment for the entire population. But what if the information *effects* of the disclosure are not randomly distributed across different groups? Disclosures A and C represent that scenario, which this Article suggests is more likely than not. Comparing Disclosure A to Disclosure C, if the goal is to maximize overall welfare, Disclosure C is the preferred choice. Most groups will increase their minimum payments, even if two groups do not. But if the goal is to increase

---

[44] *See, e.g.,* Oren Bar-Gill & Ryan Bubb, *Credit Card Pricing: The CARD Act and Beyond*, 97 CORNELL L. REV. 967, 969, 1003-05 (2012).

[45] The amount is the increase in average monthly payments made pursuant to a given type of disclosure, with red amounts signifying that the disclosures resulted in a decreased average monthly payment.

payments while also minimizing harm (reducing payments), Disclosure A is the correct choice. How then to maximize utility for all groups? In the world of digital contracting, it would be possible, instantly, to display the utility maximizing disclosure at the beginning of the formation process – to different consumer groups – yielding optimal choices.

The aforementioned example illustrate the limits of both the law and economics and BLE approaches to disclosure and contract formation. It is not just that consumers are not rational. It is not just that BLE insights can help reduce general error rates across the entire population. In fact, it may be that error rates are non-randomly distributed across groups for a variety of reasons, and if so, corrective measures require a differentiated and discriminating contract formation regime to maximize the utility of sub-groups collectively. This will allow for higher social welfare across all groups compared to a standardized approach using a single blunt disclosure instrument or formation method. And digital platforms provide both an easy test method and a cost-less ability to make contract-formation changes.

But why might decision-making errors be non-randomly distributed across certain consumer populations and how would a race-conscious contract-formation process solve for them? A few detailed examples might provide further context. First, language differences might result in formation inefficiencies. Second, socio-demographic differences might cause formation inefficiencies. Finally, differentials in consumer engagement with contract terms and disclosures might cause formation inefficiencies. Let's take each case in turn.

A. *Language Barriers*

Scholars have focused on comprehension issues with respect to disclosure, insofar as disclosures are to be designed to reflect a uniform consensus about how standard English language speakers process information. Even if one accepts as a given that disclosures are generally designed to reflect text for consumers with an 8th grade reading level, this still presumes that all consumers read English at that level – and in the same way. These assumptions are false.[46]

---

[46] Almost ¼ of the U.S. population over the age of 5 speaks a language other than English at home. *See*, *e.g.*, U.S. Census Bureau, *2016 American Community Survey 1-Year Estimates, Language Spoken At Home by Ability to Speak English for the Population 5 Years and Over* ("2016 ACS Home Language Data"), https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_B16001&prodType=table.

One real-life illustration of this fallacy shall suffice. The Consumer Financial Protection Bureau ("CFPB") thoughtfully reassessed its consumer education program with respect to Real Estate Settlement Procedures Act ("RESPA") and TILA disclosures for home-buyers, and it launched an innovative new education regime for the entire U.S. population of mortgage consumers.[47] Proactively recognizing that a large segment of U.S. home buyers spoke Spanish as a first language, the CFPB undertook the process of translating the finished English-language mortgage acquisition information and disclosure materials into Spanish. This is no small empirical feat. Literal translation – of the Google Translate variety – is not effective for the sort of sophisticated consumer education such documents are intended to convey. Further, dictionary translations across languages have at their core a false equivalency assumption – namely that standard and familiar terms can be easily translated across languages without cultural context clues.[48]

After the CFPB outsourced its English-language disclosure and information materials and translated the material for a Spanish language audience, the CFPB and its language translation team learned through small-scale focus group testing that the translations were not effective.[49] In their words, their translation team "identified particular concepts that could pose problems in the translation. These concepts did not translate directly into Spanish, did not have a definite term across multiple dialects, or the concepts behind the terms were inherently difficult. These terms included: *Appraisal, Balloon Payment, Borrower, Escrow, Final Payment, and Origination Charges*."[50]

Through extensive revision, the CFPB was able to find the appropriate language benchmarks, notwithstanding inter-cultural differences in

---

[47] *See* Alexander Bader, *Truly Protecting the Consumer in Light of the Subprime Mortgage Crisis: How Generally Applicable State Consumer Protection Laws Must Be a Key Tool in Keeping Lending Institutions Honest*, 25 J. CIV. RTS. & ECON. DEV. 767, 782-83 (2011) ("[A]n unfortunate reality of both RESPA, and its predecessor TILA, is that they had little effect on borrowers' decision-making because many mortgages are difficult for a lay person to understand on his or her own.").

[48] Consider a reverse example. The Korean idiom 똥 묻은 개가 겨 묻은 개 나무란다 literally translates in English to "A dog with feces scolds a dog with husks of grain" when in its cultural context, is meant to communicate an idea similar to the English-language idiom "People who live in glass houses shouldn't throw stones."

[49] The difficulty of having an English dominant approach to multi-lingual focus groups, the challenge of translating concepts, and the interpretation of the meaning of such concepts as tied to identity are difficult subjects for any researcher to tackle, especially the CFPB. *See, e.g.*, Taeku Lee, Language-of-Interview Effects and Latino Mass Opinion (April 2001). JOHN F. KENNEDY SCHOOL OF GOVERNMENT FACULTY RESEARCH WORKING PAPER SERIES 01-041.

[50] *See* Kleinmann Communication Group, *supra* note 39, at p. vi.

interpretation across various Spanish-speaking subgroups.[51] In 2016, the CFPB's mortgage disclosure and information regimes were finally translated to Spanish as a result of an expensive and thoughtful proactive government response. How should the remaining millions of non-native English speaking home-buyers be educated about the process of home ownership? Should government bear the considerable burden of multiple iterations and translations of standard disclosures?[52] If not, should financial institutions and other large corporations be accountable?[53] If neither should be accountable (as most present regimes contemplate), how should we expect millions of English as a second language speakers to correctly interpret disclosure materials and contract terms that they not only do not understand in English but that may, in fact, be incorrectly "translated" into false-equivalent terms through the use of basic technological translation devices that consumers might seek on their own? The solution is straightforward – allowing for consumers to have unrestricted language opt-in and requiring testing and refining of translated disclosure by regulated entities. Further, the dissemination cost (and perhaps the efficacy) of such disclosure is substantially reduced when it is deployed through digital methods, rather than burdensome traditional mailings or in-person lengthy disclosure forms.

B. *Socio-demographic Decision-Making and Behavioral Differences*

Levels of financial education, educational attainment, and the interactions of those factors with a consumer's socio-demographics may structure market choices and contract formation in complicated ways.[54] But the provision of standardized disclosure and standardized contract terms ignores these differences and assumes a uniform mono-cultural response. And, financial regulations like the Equal Credit Opportunity Act ("ECOA") impose race-gathering restrictions on various creditors and financial institutions and prohibit them from considering a consumer's

---

[51] *Id.*

[52] As of January 2019, consumer ECOA guidance brochures, for example, are only available in English and Spanish. See *Final Language Access Plan for the Consumer Financial Protection Bureau* available at https://www.federalregister.gov/documents/2017/11/16/2017-24854/final-language-access-plan-for-the-consumer-financial-protection-bureau#footnote-4-p53482 (Last visited Feb. 24, 2019).

[53] Market leaders may benefit by customer acquisition and satisfaction if they engage this effort. *See*, *e.g.*, Molly Kissler, *400,000 Chase customers opt for Spanish-language statements*, PHX. BUS. J. (Aug. 6, 2010), https://www.bizjournals.com/phoenix/stories/2010/08/02/daily72.html (describing Chase's commitment to Spanish-language access.)

[54] *See* Richard Epstein, *The Dangerous Allure of Libertarian Paternalism*, 5 REV. OF BEHAV. ECON. (2018) at 405-406.

background, even when such consideration might benefit the consumer, or at the very least, provide the consumer with data that could be used in a potential claim for effects-based discrimination.[55]

Compare the commercial-law approach to race-uniform decision-making to decision-making architecture in other areas. For instance, medical research shows disparities in the ways in which different races and genders approach medical issues.[56] Analyzing the differences in disease and treatment across different races/ethnic groups and genders has become a focus of medical research in topics ranging from lung cancer[57] to heart disease,[58] including over 200 drugs that currently have an FDA label including specific genetic recommendations – all in order to maximize not the "general health" but sub-population health.[59] Thus, health research and policy increasingly affirmatively discriminates with respect to information provided to consumers. While this approach has been met by some

---

[55] U.S. Gov't Accountability Office, *Data Limitations and the Fragmented U.S. Financial Regulatory Structure Challenge Federal Oversight and Enforcement Efforts*, CQ TRANSCRIPTIONS, LLC, 19-20 (July 15, 2009), stating that

> A final data limitation is that depository institution regulators generally do not have access to personal characteristics data (for example, race, ethnicity, and sex) for nonmortgage loans, such as business, credit card, and automobile loans. In a 2008 report, we reported that Federal Reserve Regulation B generally prohibits lenders from requesting and collecting such personal characteristic data from applicants for nonmortgage loans. . . . In the absence of personal characteristic data for nonmortgage loans, we found that agencies tended to focus their oversight activities more on mortgage lending rather than on areas such as automobile, credit card, and business lending that are also subject to fair lending law. . . . [S]uch procedures had a high potential for error and were time-consuming and costly.

[56] Anna Kline, *Pathways into Drug User Treatment: The Influence of Gender and Racial/Ethnic Identity*, 31:3 SUBSTANCE USE & MISUSE 323 (1996) (analyzing patterns of behavior in different races and genders; finding, for instance, that 'Hispanics' were more likely to delay medical treatment than other races due a discomfort or reluctance to acknowledge their addictions).

[57] Delia A. Dempsey et al., *Genetic and Pharmacokinetic Determinants of Response to Transdermal Nicotine in White, Black and Asian Non- Smokers*, CLINICAL PHARMACOLOGY & THERAPEUTICS (2013) (available at doi:10.1038/clpt.2013.159) (stating that lung cancer is typically correlated with smoking behavior, such as number of cigarettes per day and ability to quit, and such behavior is linked to the rate of metabolism of nicotine, which varies by race and ethnicity).

[58] Nicholas Wade, *Race-Based Medicine Continued...* N.Y. TIMES (Nov. 14, 2004), http://www.nytimes.com/2004/11/14/weekinreview/14nick.html?_r=0 (discussing research that indicated that the heart disease medication BiDil was more successful in treating Black patients and discussed the human genome project, which is likely to produce diagnostic tests and treatments specifically tailored to specific populations).

[59] Linda M. Hunt, Nicole D. Truesdell, & Meta J. Kreiner, *Genes, Race, and Culture in Clinical Care: Racial Profiling in the Management of Chronic Illness*, 27:2 MED. ANTHROPOLOGY Q. 253 (2013).

criticism, medical research continues to look for areas to personalize medicine – by race – in order to increase its effectiveness.[60]

C. *Differential Disengagement*

Much has been written about whether individuals actually read disclosure, concluding that they do not, we accept the insights of that literature as given.[61] Thus, a portion of the population derives no utility from disclosure. If non-readers generally make less efficient choices for contract terms, how might the existing disclosure regime be modified to induce the behavior that it seeks? Questions regarding the quality of information provided in a disclosure and the nexus of such information to being read and understood are difficult to answer and rarely asked.[62]

Some speculations follow. Perhaps a consumer might opt into interest-based financial education through a disclosure regime by an online provider.[63] The provider might be permitted to inquire about a consumer's key interests (whether sports, dance, film, etc.) Then, disclosures and explanations of key contract terms could be modified or supplemented with consumer finance scenarios that directly engaged the consumer's core interests. For example, a music fan might receive a disclosure that involved purchasing a pair of concert tickets on a credit card and explaining how the face-value of the tickets might not reflect the actual cost if the tickets were carried as credit card debt for three months at a given interest rate. Perhaps the information could use music analogies or local artists as examples to generate more consumer interest and thus increase the likelihood that the disclosure would be both accessed and understood.

With due care, lenders could also use cultural references that resonated with their audience. Thoughtful critics have suggested that tailored messages during contract formation might prove to be culturally insensitive. One response might be that, at present, the entire online

---

[60] *Id*. (stating, "[s]ome argue that taking race/ethnicity into consideration is clinically useful and can provide convenient insight into a patients' genetic heritage, behavioral habits, and socioeconomic status (citation omitted). Others argue that such practices are not scientifically defensible and may increase disparities by promoting stereotyping.")

[61] *See, e.g.,* Florencia Marotta-Wurgler, *Does Disclosure Matter?*, NYU LAW & ECON. RES. PAPER NO. 10-54 (2010).

[62] *See, e.g.,* J. H. Verkerke, *Legal Ignorance and Information-Forcing Rules*, 56 WM. & MARY L. REV. 899, 939 (2015) (evaluating information-forcing default rule research about how to make such rules effective, and finding only "a few scholars" have produced such work.)

[63] *See*, *e.g.*, Woodrow Hartzog, *Website Design as Contract*, 50 AM. U. L. REV. 1635 (2011) (providing framework for how such an approach can balance privacy and information security and be achieved using traditional contract principles).

contract formation regime is culturally insensitive, because it simply ignores race and other variables under the guise of a uniform "generic" disclosure or standard set of formation terms. In short, if consumers who are not presently engaged with the information in the disclosure were provided incentives to read the disclosure, those incentives would improve utility at no cost to those who did not receive the information.[64] Or, more creatively, disclosures could take the form of videos, snapchats, music, or other forms of communication that might more effectively reach and engage the intended audience.[65] Though this technique does *not* require the use of race variables, the methodology by which customers preference-ordered or shared information might be correlative.[66]

Therefore, with respect to the broad categories above (language access differences, socio-demographic or cultural differences, and information engagement differences), the gathering of socio-demographic information and its associated use to calibrate more efficient and effective contracts would have the net effect of enhancing overall consumer utility *and* net utility of marginalized groups. Further, corporations that excel at reducing such disparities would retain a unique marketplace advantage: proof that diverse customers of "Citilend" default on loans less frequently, demonstrate greater increases in credit scores over time, are more likely to gain access to other credit products, and other such indicators would enhance Citilend's customer base and serve to calibrate its brand identity, particularly within communities that are skeptical of large financial institutions.[67] But right now, with government controlling and mandating

---

[64] In other contexts, such as Google Ad placement, responses to inquires suggested that perceived race of the person "queried" was utilized to differentiate ads returned in the query response, presumably maximized to get a higher click-through rate. *See* Latanya Sweeney, *Discrimination in Online Ad Delivery,* DATA PRIV, LAB (2013), https://dataprivacylab.org/projects/onlineads/1071-1.pdf). Such algorithms could also be used to create, modify, or supplement disclosures in a manner consistent with the grantee's requests for more information.

[65] *See*, *e.g.*, M. Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 NOTRE DAME L. REV. 1027, 1032–34 (2012).

[66] In an alternative framework (separate from race), for example, lenders and issuers could ask a series of drop-down questions about a consumer's interests (similar to how Tivo or Netflix or Amazon fine-tune recommendations based on ratings and/or viewing/buying behavior). This preference ordering could be used to deliver extremely granular information –making it more likely to be seen and utilized by the consumer. I thank Josh Bowers for this observation.

[67] *See*, *e.g.*, Erik Oster, *Most Marketers Agree Diverse Images in Ads Help a Brand's Reputation, According to New Report*, ADWEEK (Dec. 5, 2017), available at https://www.adweek.com/brand-marketing/most-marketers-agree-diverse-images-in-ads-help-a-brands-reputation-according-to-new-report/ (explaining that in product advertising, for example, "[m]arketers are also recognizing that choosing images that are relatable to diverse groups benefits their brand's reputation."); *See also* Phil Schrader, *Why Committing to LGBT Equality and Embracing a Diverse Workplace Is So Good for Brands*, ADWEEK (April 16, 2017)

the entirety of the disclosure product in many regimes, the marketplace value (for providers *and* consumers) for better financial disclosure, among other things, is simply unknown.[68]

III. OBJECTIONS TO THE SUB-GROUP DISCRIMINATION APPROACH

A restructuring of our existing approaches to contract disclosure and formation to incorporate consumer-level sub-group differences might raise a variety of objections. First, what evidence do we have that subgroup-specific disclosures would lead to more efficient consumer behaviors than the existing disclosure models? Second, the gathering and use of this information raises online privacy concerns. Third, to successfully implement a sub-group disclosure and formation model, government must permit discrimination–or at least delineation–between certain types of sub-groups at a time when such discrimination is frowned upon in other contexts.

The first objection, lack of evidentiary proof of sub-group differential disclosure efficiency, is firmly rooted in empirics–and an absence of evidence. We know that a uniform format and dissemination model, in some contexts, works to enhance decision-making and utility for the group as a whole.[69] We do not know, as applied, whether it works the same way for population sub-groups, and there are a variety of reasons discussed above to think that it may be harmful. The best way that we can acquire objective answers to that question would be to permit or encourage a natural information experiment.[70] Those who believe that responses to disclosure and consumer errors are not randomly distributed across groups are most likely to permit or encourage a natural information experiment. Whereas those who believe that consumer errors are randomly distributed would continue to prefer the current regime. Certainly, large scale focus group testing could be conducted by corporations, by the CFPB, or by researchers. But here the focus is on real-world financial behavior and

---

available at https://www.adweek.com/brand-marketing/why-committing-to-lgbt-equality-and-embracing-a-diverse-workplace-is-so-good-for-brands/) (LGBT equality measures and sub-group centered initiatives benefit corporations by attracting talent, among other factors.)

[68] *See*, *e.g.*, Bruce R. Huber, *The Fair Market Value of Public Resources*, 103 CAL. L. REV. 1515, 1552-53 (2015) (describing valuation inefficiencies for public resources when exclusive government control of the resource obviates natural open-market pricing mechanisms).

[69] *See generally* KAZUHISA TAKEMURA, BEHAVIORAL DECISION THEORY: PSYCHOLOGICAL AND MATHEMATICAL DESCRIPTIONS OF HUMAN CHOICE BEHAVIOR (2014).

[70] *See generally* Jimenez et. al., *supra* note 10. Of course, researchers in this space can conduct focus groups and surveys, which are valid measurement tools and would inform this discussion. But here one should be particularly concerned with measuring real outcomes under real conditions–and the most critical experimental tool is thus the changing of disclosure–in context.

outcomes for marginalized groups in the United States–not merely opinions about a consumer's hypothetical behavior over time–and so any conclusions drawn from such field or one-off experiments would be necessarily limited for this reason.[71]

But the lack of empirics may exacerbate the problem. For example, "information on consumer race and ethnicity is required to conduct fair lending analysis of non-mortgage credit products, but auto lenders and other non-mortgage lenders are generally not allowed to collect consumers' demographic information. As a result, substitute, or "proxy" information is utilized to fill in information about consumers' demographic characteristics."[72] And these proxies are quite imprecise.

The second objection, that encouraging individuals to further identify race and other socio-demographic factors in online contracting may implicate privacy concerns, prompts a few responses.[73] First, failing to ask about socio-demographic factors may signal government's disinterest or communicate that government thinks race, in this setting, is not important. Second, such that permitting the use of such variables enables consumers to promote self-realization or positive identity construction, the ability to self-identify race and other factors in consumer contract regimes subject to certain restrictions can serve to enhance, not undermine, individual interests.[74] Third, the nexus of socio-demographics to privacy in the digital era is less clear than one might expect, given that privacy law scholars have not typically engaged race and other socio-demographics as key data-

---

[71] *But see* Marianne Bertrand, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, & Jonathan Zinman, *What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment*, 125 Q. J. ECON. 263, 263 (2010) (discussing a South African experiment demonstrating that consumers responded differentially to a loan-advertisement's experimentally varied terms and content). However, the path-breaking study was necessarily limited to a specific context: "mailers were sent exclusively to clients who successfully repaid prior loans from the Lender. Most had been to a branch within the past year and hence were familiar with the loan product, the transaction process, the branch's staff and general environment, and the fact that loan uses are unrestricted." Further, the study's exploration of the nexus of cultural or racial cues to response rates was inconclusive, "Given our lack of strong priors on how any advertising content effects might vary with consumer characteristics, and statistical power issues, we will not devote much space to discussing heterogeneity in responses to advertising content."

[72] *See* Consumer Fin. Prot. Bureau, *Proxy Methdology Report* 3 (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

[73] *See* Shmuel I. Becher et. al., *supra* note 29, at 31 (acknowledging same with respect to tailoring proposals to low income consumers); *See also* Woodrow Hartzog and Frederic Stutzman, *The Case for Online Obscurity*, 101 CAL. L. REV. 1, 1-2 (2013).

[74] *See, e.g.,* Jonathan D. Kahn, *Privacy as a Legal Principle of Identity Maintenance*, 33:2 SETON HALL L. REV. 371, 373-74 (2003).

points warranting additional scrutiny.[75] And, such that socio-demographic data obtained through this process is abused or misused, government can design remedies designed to penalize merchants who violate a consumer's and the government's expectations about the sharing, use, or misuse of this particularly personal information.[76] Alternatively, companies choosing to gather and use such data for these purposes could be encouraged to provide for warnings or disclosures or "demographic opt-outs," which could increase transparency and salience for consumers, allowing them to choose a more generic approach if so desired.[77]

The third objection is a moral one, that encouraging the use of socio-demographics in this way encourages invidious discrimination. For those opposed to this approach, the best way to reduce the likelihood of a discriminatory market outcome with respect to racial and other socio-demographic sub-groups in the consumer finance disclosure regime would be to prevent market-actors, government, and digital platforms – from permitting – and encouraging – discrimination on the basis of race.[78] But a response might be that we expect the government and private markets to encourage 'positive' discrimination in other contexts – such as affirmative action – where the utility of such discrimination may not be equally enhancing for all groups or may be perceived by some to be more harmful to other groups. Even in these other contexts, the gathering and analysis of "consumer" data serves as a core component of the analysis of the effectiveness of such programs.[79]

With respect to online contracts in markets that are heavily regulated by government, such as those governing financial services, mandated information gathering about race, ethnicity, and other factors is critical for

---

[75] *See*, *e.g.*, Will Thomas DeVries, *Protecting Privacy in the Digital Age*, 18 BERKELEY TECH. L.J. 283, 305-11 (2003) (surveying the array of concerns as the U.S. transitioned into a more digitally connected era, suggesting changes to traditional privacy law to modernize its focus, but not discussing socio-demographic information as an area of concern.)

[76] *See* Ari Ezra Waldman, *Privacy as Trust: Sharing Personal Information in a Networked World*, 69 U. MIAMI L. REV. 559, 628 (2015) (reframing privacy debate as one centered on trust between the sharer and the recipient and identifying a framework valuing "the socially beneficial effects of sharing and [giving] judges a coherent scheme for answering limited privacy questions.")

[77] *See, e.g.*, Gerhard Wagner and Horst Eidenmüller, *Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions*, 86:2 U. CHI. L. REV. 581 (2019).

[78] *See Parents Involved in Community Schools vs. Seattle School District #1*, 551 U.S. 701 (2007).

[79] *See, e.g.*, Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of Affirmative Action*, 94 CAL. L. REV. 1063, 1065-66 (2006) (exploring nexus of social science research about implicit bias and the effectiveness of affirmative action programs as solutions for discrimination).

supporting the type of anti-discrimination lawsuits that form the core of civil rights litigation in a variety of contexts, including acquisition of credit. For example, in the markets for consumer loans and home mortgages, lenders subject to the Housing Mortgage Disclosure Act ("HMDA") and/or Regulations B[80] and C have long been subject to rigorous data-gathering requirements, including asking online and in-person borrowers about race and other criteria. On an aggregate level, civil rights advocates and federal government researchers and law enforcers have been able to utilize this demographic information in statistical models, identify disparities across institutions, and then sue to recover damages and to eliminate racially discriminatory practices.

Absent the gathering of such data as transactions become more digitized, it is extraordinarily difficult to prove, for example, disparate impact claims. The data analyses underlying those claims must use proxies for race, ethnicity, and gender, since the variables themselves are not collected. Thus, the failure to gather (whether secretly through online tracking or openly by asking) race and ethnicity information can unintentionally benefit lenders and financial institutions, because these proxy methodologies used by regulators are imperfect and tend to overstate disparities, thus allowing lenders and others to call their validity into question.[81] It is not terribly hard to find weaknesses in the proxy measures. For example, the CFPB utilized Census track surname data from *Census 2000* to construct its associated consumer contract race and ethnicity measures – until *April 2017*.[82]

---

[80] Reg. B institutions that receive "an application for credit primarily for the purchase or refinancing of a dwelling occupied or to be occupied by the applicant as a principal residence, where the extension of credit will be secured by the dwelling, shall request as part of the application" the marital status, age, ethnicity, race, and gender of the applicant. Historically Reg B data included five data fields: ethnicity, race, sex, marital status, and age, while HMDA included only ethnicity, race, and sex. Regardless, these socio-demographic variables could be used in a statistical analysis in order to test for their effects holding constant other factors, like credit scores.

[81] *See, e.g.*, Am. Fin. Serv. Ass'n, *Request for Information Regarding the Bureau's Inherited Regulations and Inherited Rulemaking Authorities* 2 (June 25, 2018), https://www.afsaonline.org/portals/0/Legal%20and%20Reg/AFSA%20-%20RFI%20on%20Inherited%20Rules%20-%20June%2025%202018.pdf (critiquing the ability of plaintiffs to prove disparate impact for a variety of reasons, including the lack of self-disclosed individual level data). *But see* D Adjaye-Gbewonyo et al., *Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study*, 49(1) HEALTH SERV. RES. 268, 277-81 (2014) (concluding the BISG method [which is the CFPB's preferred] may indeed be useful for classifying race/ethnicity of health plan members when needed for health care studies).

[82] *See, e.g.*, *Update to Proxy Methodology,* GITHUB (Apr. 2017), https://github.com/cfpb/proxy-methodology.

Now imagine how such race proxies are operationalized when it comes to that have become more digitized over time; this digitization does not necessarily expand anti-discrimination norms nor follow a Bentham-like utility maximizing path. Consider the case of the CFPB's now-rescinded indirect auto-lending discrimination rules. Indirect auto-lenders were subject to the same ECOA restrictions as credit-card companies–namely they were forbidden from gathering certain socio-demographic information, including race, ethnicity, and sex.[83] As a result, it was difficult to test for disparate impact or discrimination in the auto-lending market, because such information was not available. However, indirect auto-lenders engaged in lending practices that yielded differential effects by racial sub-groups, through practices such as differential mark-ups on the "dealer reserve".[84] To solve for the information-analytics gap, civil rights advocates lobbied the CFPB and others to allow for the use of "race proxies" in the data-analysis process, because neither in-person nor online transactions permitted its acquisition.

But this use of such proxies gave indirect auto-lenders an easy rhetorical target–a flawed methodology would lead to industry ruin. When indirect auto-lenders saw that the CFPB intended to subject them to ECOA scrutiny for alleged discriminatory pricing racial disparities using this flawed methodology, they lobbied Congress to overturn the regulation.[85] And, though not solely for that reason, Congress agreed. The indirect auto-lending regulation was overturned on May 21, 2018.[86] Now the leading online direct/indirect auto-lending markets still lack broad-based race-data and may have discriminatory racial impacts, but there is still no way to directly test for such impacts or to provide consumers with a socio-demographically attuned contract model. This same lack of supply-side socio-demographic information from those searching for online pay-day loans also undermines efforts to prove that pay-day lending contracts are discriminatory.[87]

---

[83] *See* Cons. Fin. Prot. Bureau, *CFPB Bulletin 2013-2* (2013), https://files.consumerfinance.gov/f/201303_cfpb_march_-Auto-Finance-Bulletin.pdf (describing the CFPB's then-interpretation of ECOA with respect to indirect auto-lenders).

[84] They were able to mask discriminatory behavior, in part, due to the *absence* of race data associated with each consumer contract.

[85] *See, e.g.*, Daniel Goldstein, *Car Dealers Win First Round in Congress Against CFPB Over Auto Loan Discrimination*, MARKETWATCH (Aug. 1, 2015), https://www.marketwatch.com/story/did-congress-just-make-it-easier-for-auto-dealers-to-discriminate-against-you-2015-08-0*1*.

[86] *See* Cons. Fin. Prot. Bureau, *Indirect Auto Lending and Compliance with the Equal Credit Opportunity Act* (Mar. 21, 2013), https://files.consumerfinance.gov/f/201303_cfpb_march_-Auto-Finance-Bulletin.pdf.

[87] *See*, *e.g.*, Paige Marta Skiba, *Regulation of Payday Loans: Misguided?*, 69 WASH. & LEE L. REV. 1023, 1038-41 (2012).

CONCLUSION

Scholars have improved both traditional and digital contract disclosure and formation models by incorporating behavioral insights in an effort to improve choice architecture and enhance overall utility. But these efforts fail to engage the utility of tracking the effect of race on contracting norms, selection of contracting terms, and contract literacy/engagement. To encourage the gathering (and the analysis) of such information, particularly when government has a role in the contract-acquisition and formation process between merchants and consumers, government must mandate that online disclosures and consumer contracts seek race and other demographic information. It is clear that the present incarnation of the CFPB, which could lead this effort, seems to have substantially downgraded the importance of analyzing race, sex, and other socio-demographic factors related to consumer information provision and consumer anti-discrimination principles in both print and digital spaces.[88]

We are just at the beginning stages of understanding how race matters in consumer disclosure and consumer contracts. But, if gathering and experimenting with such information allows for online platforms to customize interfaces and disclosures such that they more effectively reach and engage diverse consumers, then there may be an increase consumer utility. Further, a more transparent gathering and use of such data will allow government (and private actors) to better maintain and enforce anti-discrimination principles because they can monitor outcomes in ways that are presently deeply imperfect. Although this race-conscious approach can be operationalized across many sectors, its value may be most salient in digital transactions in regulated industries, such as consumer finance, where government has the greatest interest in evaluating market engagement of consumers of various racial backgrounds to ensure fairness.

---

[88] *See, e.g.*, Cons. Fin. Prot. Bureau, *Summer 2018: Supervisory Highlights,* 12 (Sept. 2018), https://files.consumerfinance.gov/f/documents/bcfp_supervisory-highlights_issue-17_2018-09.pdf (discussing none of these subjects, absent a single sentence).

ARTICLE

EXPLANATION < JUSTIFICATION: GDPR AND THE PERILS OF PRIVACY

TALIA B. GILLIS AND JOSH SIMONS[†]

The European Union's General Data Protection Regulation (GDPR) is the most comprehensive legislation yet enacted to govern algorithmic decision-making. Its reception has been dominated by a debate about whether it contains an individual right to an explanation of algorithmic decision-making. We argue that this debate is misguided in both the concepts it invokes and in its broader vision of accountability in modern democracies. It is justification that should guide approaches to governing algorithmic decision-making, not simply explanation. The form of justification – who is justifying what to whom – should determine the appropriate form of explanation. This suggests a sharper focus on systemic accountability, rather than technical explanations of models to isolated, rights-bearing individuals. We argue that the debate about the governance of algorithmic decision-making is hampered by its excessive focus on privacy. Moving beyond the privacy frame allows us to focus on institutions rather than individuals and on decision-making systems rather than the inner workings of algorithms. Future regulatory provisions should develop mechanisms within modern democracies to secure systemic accountability over time in the governance of algorithmic decision-making systems.

[†] Talia Gillis is an Empirical Law and Finance Fellow and SJD candidate at Harvard Law School, and a PhD candidate in Business Economics. Josh Simons is a PhD candidate in Government at Harvard University. He is a Graduate Fellow at the Edmond J. Safra Centre for Ethics and an Affiliate at the Berkman Klein Centre for Internet and Society.

INTRODUCTION

The European Union's General Data Protection Regulation (GDPR) is the most comprehensive legislation yet enacted to govern algorithmic decision-making. Its scope is supra-national, shaping the data protection practices of companies operating throughout the world's most prosperous integrated economic area. It establishes enforcement mechanisms with bite, threatening companies with fines of up to 4 percent of global turnover for the most serious violations. Yet the GDPR's focus is not decision-making, but privacy. This is the product of history. The primary protagonists of current debates about governing algorithmic decision-making are privacy scholars. We believe this privacy lens has distorted interpretations of the GDPR's approach to governing algorithmic decision-making. That approach reaches beyond an individual right to explanation, to establish provisions that aim to build systemic accountability over time.

This paper examines those provisions. We explore the tools the GDPR provides for ensuring that institutions justify their use of algorithmic decision-making systems, to both regulators and individuals subject to their decisions. Our aim is not simply to interpret the GDPR, though we side with scholars who argue that the main text of the GDPR must be read in conjunction with surrounding 'soft-law', including the Recitals, Article 29 Working Party (A29WP) guidance, and the interpretations of authorities mandated with enforcing its provisions.[1] Rather, our aim is to step back and examine the concepts that underpin the right to explanation debate, and the broader challenge of regulating algorithmic decision-making. We make three arguments.

---

[1] *See* Margot E. Kaminski, *Binary Governance: Lessons From the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. (forthcoming 2019), at 48; Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 197-199 (2019); Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143 (2019).

First, we argue that accountability is the foundational goal that should guide approaches to governing algorithmic decision-making. Accountability is achieved when an institution must justify its choices about how it developed and implemented its decision-making procedure, including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and enforcement. Accountability produces instrumental benefits, including encouraging the use of decision-making procedures that are consistent and verifiable, and providing mechanisms for identifying and addressing discrimination and injustice.[2] However, we argue that accountability is the foundational goal because of its intrinsic, rather than its instrumental value. Accountability is constitutive of democratic self-governance. It is part of what it means for a citizenry to authorize in an ongoing way the complex decision-making systems whose recommendations shape their lives. Other goals discussed in the literature are all in some way means to securing accountability. Individual explanations of algorithmic systems are valuable if and when they enable institutions to justify those systems to individuals and regulators, but they may not always further this end.[3] Transparency may be necessary for some forms of accountability, but neither constitutes nor is entirely sufficient for accountability.[4] In other words, accountability requires justification and justification requires explanation. The form of each should determine the form of the others.

Second, we distinguish between different forms of justification required to attain systemic accountability and consider the appropriate form of explanation in each. Recent scholarship has debated whether a right to an explanation exists in the GDPR, and if so, what its content might be. We argue that the form this explanation should take must depend on the form of accountability being pursued. By separating out different forms of justification, we set out how a 'right to explanation' (a "RtE") might further the aim of accountability, and how it might

---

[2] Jeremy Waldron, *Accountability: Fundamental to Democracy* 26 (NYU Pub. Law & Legal Theory Research Paper Series, Working Paper No. 14-13, Apr. 2014), https://papers.ssrn.com/abstract=2410812; MATTHEW V. FLINDERS, THE POLITICS OF ACCOUNTABILITY IN THE MODERN STATE (Aldershot: Ashgate 2001); ADAM PRZEWORSKI, SUSAN CAROL STOKES, AND BERNARD MANIN, DEMOCRACY, ACCOUNTABILITY, AND REPRESENTATION (Cambridge Univ. Press 1999).

[3] *See, e.g.,* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

[4] *See generally* Mike Ananny & Kate Crawford, *Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973 (2016); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1530 (2013).

hinder it. We argue that the technical explanation of a statistical or machine learning model is not sufficient for an institution to justify its decision-making procedure. Furthermore, we argue that such a technical explanation may even distract from the most important provisions of the GDPR for securing systemic accountability.[5] These include two crucial components. First, a range of mechanisms for ensuring that institutions justify their choices in the design and implementation of algorithmic decision-making systems – the critical *ex ante* stage in machine learning – including their broader policy and commercial aims. Second, that these mechanisms ensure justifications are offered to regulators with the necessary information, resources and powers, not simply isolated, rights-bearing individuals with limited information and expertise.

Third, we argue that the GDPR's focus on privacy underpins some of its most significant limits. We identify three such limits, some of which are about the law itself, others about recent interpretations of the law. First, recent interpretations of the law mistakenly focus on the actual algorithms and machine learning models, rather than the broader policy and commercial environment in which they are deployed. The aims an institution has in designing and implementing an algorithmic decision-making system shape the workings of the algorithm or model itself, but receive far less attention, at least in the interpretive literature. Second, the law itself is constrained by its focus on individual rights. Machine learning, the most common form of algorithmic decision-making, makes information about the design and implementation about the overall system critical to exercising meaningful oversight. Information about individual decisions will not enable individuals to grasp of the nature of the system whose decisions shape their lives, or enhance their capacity to demand a justification from the powerful institutions that designed it. For related reasons, the notice and consent framework is not an adequate mechanism by which to ensure meaningful institutional accountability. Third, the GDPR and the literature surrounding it have no satisfactory account of how its provisions are to be subject to democratic oversight. Accountability matters because it is constitutive of democratic self-government. Future regulatory provisions must focus more directly on developing mechanisms within modern democracies that can secure accountability in the governance of algorithmic decision-making systems.

---

[5] *See, e.g.,* Lillian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 19, 65-67 (2017); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SEC. & PRIV. 46, 50 (2018).

The paper proceeds in two sections. The first contains our conceptual argument. We begin by arguing that accountability is the foundational goal that should guide approaches to governing algorithmic decision-making. Explanations are instrumentally valuable insofar as they enable the process of giving and receiving justifications that constitutes accountability in a democracy. The second draws out the implications of this argument for interpreting the GDPR and for approaches to governing algorithmic decision-making more broadly. We focus specifically on machine learning in this paper. Though we are interested more broadly in governance approaches to algorithmic decision-making, focusing specifically on machine learning draws attention to the most acute practical and theoretical challenges. We focus our discussion on governing the use of machine learning in the private rather than the public sector.

We end by setting out some of the ways in which the limits to recent interpretations of the GDPR are related to their framing in terms of privacy. The challenges we face when developing governance systems for algorithmic decision-making go beyond concerns that can usefully be understood in terms of privacy.

I. EXPLANATION → JUSTIFICATION → ACCOUNTABILITY

This section outlines our conceptual argument. First, we argue that accountability is the foundational goal. It should guide our interpretations of the GDPR. It should also drive the questions we pose, and the answers we advance, as we confront the broader task of developing a comprehensive approach to governing algorithmic decision-making. Second, we consider the implications this has for the other concepts invoked in the debate about whether a right to explanation exists in the GDPR (hereafter the RtE debate). The most obvious is explanation itself, providing explanations of the logic of a machine learning model to ensure that its operation is, in some way, comprehensible to external human observers. Explanations are said to be valuable because there is something inherently important about individuals understanding the systems to which they are subject, that is, because they respect individual autonomy; and also because such understanding is instrumentally important, for individuals to challenge decisions or to identify bias and discrimination.[6] We argue that

---

[6] *See* Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 243, 250 (2017); Selbst & Barocas, *supra* note 3, at 40-46.

explanations of machine learning models are valuable if and when they are a means to provide justifications of the broader decision-making procedure. What matters is justifying why the rules are the way they are; explaining what the rules are must further this end.

The focus on individual, technical explanations has been driven by an uncritical bent towards transparency. Transparency is thought to matter because to see is to know, and knowledge is power. Transparency provides the information required for governance and oversight.[7] This is a mistake. Like explanation, transparency is an instrumental good. Transparency matters if and when it is required to further the aim of systemic accountability. These concepts are important not only for the RtE debate, but for thinking more broadly about the central aims that should guide any approach to governing algorithmic decision-making. This section aims to make progress towards such conceptual clarity.

A. *Accountability*

Accountability, we submit, is the foundational concept. It is the motivation that drives arguments for transparency and for various forms of explanation in machine learning. It should be the central aim of all approaches to governing decision-making using machine learning. It is therefore important to be clear about what accountability is and why it is valuable.

Accountability is about vertical power. Accountability empowers those who might otherwise be powerless, demanding that those who wield power over them offer an account of their conduct. In the modern world, its most familiar form is democratic accountability, in which those who control the apparatus of well-organized territorial states must offer an account of their conduct to citizens subject to their power. Democratic accountability, as Jeremy Waldron puts it, confers "authority on those who are otherwise powerless over those who are well endowed with power."[8] More generally, accountability can be said to pertain in the following structure. Party A is accountable to party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate.[9]

---

[7] *See* Ananny & Crawford, *supra* note 4, at 974-977; Zarsky, *supra* note 4, at 1533; *See generally,* DAVID BRIN, THE TRANSPARENT SOCIETY: WILL TECHNOLOGY FORCE US TO CHOOSE BETWEEN PRIVACY AND FREEDOM? (1998).

[8] *See* Waldron, *supra* note 2.

[9] *See* Reuben Binns, *Algorithmic Accountability and Public Reason*, 31 PHIL. & TECH. 543, 544 (2018); *see generally* MARK BOVENS ET AL., THE OXFORD HANDBOOK OF PUBLIC ACCOUNTABILITY (2014).

This is the principal-agent of accountability.[10] Accountability requires an agent, such as rulers, to justify their conduct to a principal, such as an electorate, subject to sanction through a range of mechanisms, most obviously, elections. The agent's exercise of power is shaped by the knowledge of the principal's inevitable judgement. Accountability ensures that those with power must justify their decisions to those who they will affect. Much like the threat of punishment, the idea is that this will change the behaviour of those decision-makers for the better.[11] To apply this view of accountability to decision-making procedures that use machine learning, let us suppose accountability pertains when: An institution (Party A) must justify its choices about how it developed and implemented its decision-making procedure (Conduct C), including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and enforcement (Party B).

Accountability can secure a range of instrumental benefits. It encourages institutions to use decision-making procedures that are consistent and verifiable, as consistency and verifiability tend to make for more persuasive justifications. It encourages institutions to identify discrimination in their decision-making procedures, and where possible, to address it in the design stage.[12] Structures of accountability can incentivize institutions to develop decision-making procedures with more care, consider a broad range of interests and perspectives, and evaluate more kinds of risk and possible harms.[13]

But accountability is about more than power. Part of the value of accountability is that it changes the conduct of those with power because they know that conduct will have to be justified. However, the more fundamental value of accountability is intrinsic. It is constitutive of democratic self-governance.[14] A king might fear the judgement of his

---

[10] This has been the dominant view of accountability explored in political science and political economy for the past two decades. *See generally* PRZEWORSKI ET AL., *supra* note 2; James D. Fearon, *Self-Enforcing Democracy*, 126 Q.J. ECON. 1661 (2011); FLINDERS, *supra* note 2; KAARE STRØM, WOLFGANG C. MÜLLER, AND TORBJÖRN BERGMAN, DELEGATION AND ACCOUNTABILITY IN PARLIAMENTARY DEMOCRACIES (2003).

[11] ROBERT D. BEHN, RETHINKING DEMOCRATIC ACCOUNTABILITY 3 (2001).

[12] Selbst & Barocas, *supra* note 3, at 42, 55.

[13] *See* Binns, *supra* note 9, at 547; Zarsky, *supra* note 4, at 1530-1550.

[14] This is part of Jeremy Waldron's argument, drawing on several recent critiques of the narrowness of the principal-agent approach to accountability, and considering its relationship to democracy more broadly. Waldron, *supra* note 2. *See also* JOHAN P. OLSEN, DEMOCRATIC ACCOUNTABILITY, POLITICAL ORDER, AND CHANGE: EXPLORING ACCOUNTABILITY PROCESSES IN AN ERA OF EUROPEAN TRANSFORMATION (2017); CRAIG T. BOROWIAK, ACCOUNTABILITY & DEMOCRACY: THE PITFALLS AND PROMISE OF POPULAR CONTROL (2011); Alexander H. Trechsel, *Reflexive Accountability and Direct Democracy*, 33 W. EUR. POL. 1050 (2010).

subjects. He might fear rebellion or resistance. The anticipation of that rebellion or resistance might shape the decisions he makes. But this is not accountability. The King need not justify his decisions; he has no obligation to offer an account of the decisions he has made, or his reasons for making them, to his subjects. Whereas in a democracy, as Waldron argues, "the accountable agents of the people owe the people an account of what they have been doing, and a refusal to provide this is simple insolence."[15]

Accountability is part of the practice of modern democracy. The giving and receiving of justifications is part of what it means to jointly govern ourselves. The agents who give and receive justifications are varied: sometimes individual citizens justify what they do or decide to other individual citizens, sometimes institutions justify what they do or decide to individual citizens, sometimes institutions justify what they do or decide to other institutions.[16] The content of their justifications might be varied too, including important decision-making processes and procedures that shape the lives of citizens. This broader view of accountability extends beyond the public realm. The most obvious form of accountability in a democracy is certainly the justification by public bodies of their conduct to citizens. But the rules and procedures that shape our collective future go far beyond those authored in the public realm. We expect companies who deliver important services to justify their decisions and procedures, to us as citizens, and to governments as our representatives. Facebook must justify how it moderates content to Congress.[17] Its content moderation system profoundly shapes how we interact as citizens; decisions about how that system works are of public concern; therefore, Facebook must justify those decisions to us, the public, or to our representatives.

Democracy and accountability are not, however, the same thing.[18] There may actually be important tensions between democracy and accountability. Mechanisms for accountability are often solutions to the problem of control – they need not, and often are not, democratic. Central banks and financial regulators are institutions of accountability, that is,

---

[15] Waldron, *supra* note 2, at 28.

[16] We side with Waldron on this point: whether the justification is offered or received by an individual, a multitude, or a legal corporation doesn't matter as much as some suppose. *Id.*

[17] *See* Kate Klonick, *Facebook Released Its Content Moderation Rules. Now What?*, N.Y. TIMES (April 26, 2018), https://www.nytimes.com/2018/04/26/opinion/facebook-content-moderation-rules.html.

[18] Mechanisms of accountability may actually change how we do democracy. If accountability changes democracy, the two cannot be synonymous. *See generally* OLSEN, *supra* note 14; Trechsel, *supra* note 14.

they solve the problem of control, but they are not democratic. Environmental regulators are institutions of accountability, but they are not democratic. As Borowiak puts it, "accountability institutions can create veils of legitimacy that mask abuses and dampen the critical and participatory energies of the public. So doing, they can end up thwarting citizen control rather than enhancing it."[19] These are important issues, but we want to put them aside. The challenge of ensuring that institutions of accountability do not erode the possibilities of democratic action and legitimacy is critical to the future of democracy in increasingly complex societies and economies, but it is a separate challenge to thinking systematically about accountability and what is required to achieve it.

Accountability, then, is the foundational concept. What follows, we believe, is that transparency must be put in its proper place. Transparency is valuable insofar as it furthers the aim of accountability.[20] The conditions in which transparency furthers this aim are more limited than it is often supposed. Demands for transparency tend to assume that if people are provided with the necessary information, they will take action against decisions they think are wrong. The GDPR, for example, requires individuals to be provided with "meaningful information about the logic involved" in the automated decision[21] as part of the right to contest these decisions[22] and to enforce other rights under the GDPR.[23]

There are good reasons to be deeply skeptical about the connection between the provision of information to individuals and those individuals taking desired actions. Firstly, people have to understand the information they receive. There is ample evidence that people struggle with even simple and straightforward disclosures,[24] let alone disclosure that pertains to more technical aspects of automated decision-making. Second, people

---

[19] *See* BOROWIAK, *supra* note 14, at 179.

[20] *See, e.g.,* Ananny and Crawford, *supra* note 4; Adrian Weller, *Challenges for Transparency*, CORNELL UNIV. (July 29, 2017), http://arxiv.org/abs/1708.01870; Danielle Citron, *What to Do about the Emerging Threat of Censorship Creep on the Internet,* CATO INST. (November 28, 2017), https://www.cato.org/publications/policy-analysis/what-do-about-emerging-threat-censorship-creep-internet.

[21] Namely Article 13(2)(f), Article 14(2)(g) and Article 15(1)(g). Regulation (EU) 2016/679 of the European Parliament and the Council of April 27, 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[22] GDPR art. 22.

[23] *See* discussion in Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT. DATA PRIV. L. 233 (2017) (showing the connection between providing information and individuals enforcing their rights).

[24] *See* discussion in Talia B. Gillis, *Putting Disclosure to the Test: Toward Better Evidence Based Policy*, 28 LOY. CONSUMER L. REV. 31, 50 (2015).

must understand how that information relates to their particular circumstances and preferences. Many years of research on the effect of information disclosure, in multiple realms, demonstrate that there is a significant gap between the promise of disclosures and their actual impact.[25] The drive towards transparency often produces legal and policy regimes that fail to achieve genuine accountability over time.

Accountability should be the central aim of all approaches to governing decision-making using machine learning. The giving and receiving of justifications is part of what it means for a citizenry to authorise in an ongoing way the complex decision-making systems whose recommendations shape their lives.

B. *Explanation < Justification*

We now turn to the central concept in the RtE debate, on which most interpretations of the GDPR have focused: explanation. On the face of it, the role of explanation in our notion of accountability seems obvious. Accountability requires justification and justification requires explanation. To justify a decision-making procedure that involves or is constituted by a machine learning model, an individual subject to that decision-making procedure requires an explanation of how the machine learning model works. This is the thought that underpins much of the RtE debate.

But let's pause for a moment to ask: What form of explanation does justification require? Think of an example. Suppose you are involved in a major car crash that leaves you paralyzed from the waist down. After you wake up in hospital, you ask: Why did I crash? The crash investigator helpfully left a report by the side of your bed. It explains: The velocity of your car produced a centrifugal force on your wheel hub, which, gradually produced a rotating motion on your wheel stud which, in turn, loosened your front left wheel from your chassis. The resulting force made your vehicle swerve to the left. The particles of the central barrier then came into contact

---

[25] *See generally* ARCHON FUNG ET AL., FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2007); Lauren Willis, *The Consumer Financial Protection Bureau and the Quest for Consumer Comprehension*, 3 RUSSELL SAGE FOUND. J. SOC. SCI. 74 (2017); OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2014); Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647 (2011); Ryan Bubb, *TMI? Why the Optimal Architecture of Disclosure Remains TBD*, 113 MICH. L. REV. 1021 (2015); Matthew A. Edwards, *Empirical and Behavioral Critiques of Mandatory Disclosure: Socio-Economics and the Quest for Truth in Lending*, 14 CORNELL J. L. & PUB. POL'Y 199, 229 (2004). For further analysis of the ideology underlying calls for transparency, s*ee* David E. Pozen, *Transparency's Ideological Drift*, 128 YALE L.J. 100 (2018) (arguing that transparency has shifted from an idea that promotes a stronger and more egalitarian regulatory state, to a tool aimed at limiting government intervention and regulation).

with the polymers on the left side of your vehicle. The molecular structure of the polymer was broken on the driver's side, rapidly reducing the speed of your vehicle and eventually bringing it to a halt. This explanation is clearly unsatisfactory. It's an explanation at the wrong level. It answers your 'why' question with an account of microphysics. You want to know why your wheel came off. The explanation might be true, but it is beside the point. What you *really* want is for Ford to justify why your wheel came off despite having serviced your vehicle last month. What matters is the justification that is part of the process of accountability. The form of explanation involved in that justification depends on the context. If Ford sent you an account of the microphysics of your crash, you would consider that not just a misunderstanding about the information you require, but an evasion of accountability. It represents a failure to justify what happened.

The RtE debate often conceives of explanations at completely the wrong level. More precisely, at a level that is simply not relevant to justification, and therefore, to accountability. To those subject to the decisions of a machine learning model, offering an explanation of a machine learning model is a little like offering an account of microphysics to explain a car crash. Explanations of machine learning models are certainly not sufficient for many of the most important forms of justification in modern democracies, and often, they may not even be necessary. More specifically, what form of explanation is necessary, including whether a technical explanation of the machine learning model is necessary, depends on who is justifying what to whom. This implies two important shifts in focus. First, in terms of *what* is being justified. The focus should be on how institutions justify their choices about the design and integration of machine learning models into their decision-making systems, rather than on what the rules governing a model's operation are. What matters is why the rules of an algorithm are what they are.[26] Second, in terms of *to whom* the justification is being offered. Institutions should justify their choices about the design and integration of machine learning models not to individuals, but to empowered regulators or other forms of public oversight bodies. Less emphasis should be placed on the rights of disempowered and isolated individuals, who are expected to understand the complicated models to whose decisions they are subject, and more on systemic accountability – the way power is structured between institutions. If accountability is the foundational goal, what is required is institutional justification, not algorithmic explanation. Algorithmic explanation can be necessary to institutional justification. But since it is justification that is necessary for accountability, and it is accountability that is of ultimate importance, the

---

[26] Selbst & Barocas, *supra* note 3.

form of institutional justification should determine the appropriate form of explanation.

The excessive focus on technical forms of explanation is itself the result of an uncritical bent towards transparency. This is partly the product of history. Much of the policy and legal debate about the governance of machine learning has developed from older debates about privacy. Many scholars who were previously privacy experts now write about the governance of artificial intelligence. The GDPR is framed as a privacy law, even though its focus reaches far beyond the confines of privacy.[27] The transparency bent, with all its pitfalls, has been unreflectively transposed from the privacy literature to the literature on explanation and interpretability.[28] The risk is that the limits of the transparency debate swiftly become limits to the debate about how we should integrate machine learning models into some of our most important social, economic and political institutions. The most important limit is the focus on individual rights, rather than on structures of power. The privacy debate has always been hemmed in by its focus on individual consent, a concept that has proved to be a mirage in theory and in practice.[29] As a result, it has overlooked more fundamental and intractable challenges about how institutions should hold one another to account, most notably, involving questions about the structure and distribution of power. If individual-understanding-of-machine-learning-models becomes the new individual-consent-to-the-use-of-their-data, we should expect a wholesale failure to hold to account the institutions that use machine learning for their own ends.

This uncritical bent towards transparency, and the subsequent focus on technical explanation, actually *suits* many of the most powerful actors in the internet age. The focus on algorithmic explanation can deflect from the need for institutional justification. Consider an example. To satisfy increasing calls for oversight and accountability in content moderation, suppose Facebook rolls out a new interactive tool. This tool allows individual users to interact with their News Feed, to understand the factors that 'explain' why they see particular pieces of content. Users would be able to change important parameters about themselves, such as their gender, race, or

---

[27] For instance, much of the literature about explanation in law is published in journals that are putatively about privacy. *See generally* Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5.

[28] *See generally* BRIN, *supra* note 7; Will Thomas Devries, *Protecting Privacy in the Digital Age*, 18 BERKELEY TECH. L.J. 283, 283–311 (2003); Joshua A. T. Fairfield & Christoph Engel, *Privacy as a Public Good*, 65 DUKE L.J. 385, 385–457 (2016).

[29] Lothar Determann, *Social Media Privacy: A Dozen Myths and Facts*, 16 STAN. TECH. L. REV. 1, 7-10 (2012); Dan Svirsky, *Why Are Privacy Preferences Inconsistent?* 15 (JOHN M. OLIN CTR. FOR LAW, ECON., & BUS. FELLOWS' DISCUSSION PAPER SERIES, HARV. LAW SCH., Discussion Paper No. 81, 2018).

location, but also their behaviour on Facebook, such as what groups they have liked, or what publishers they read, and see how their News Feed changes. No doubt many users would feel Facebook had discharged its responsibility to explain how News Feed works. But this is not satisfactory. To the question "Why do I see what I see?", the tool effectively says "Well, if you were African American and not white, you'd see this; if you were female and not male, you'd see this; if you were from California and not Wisconsin, you'd see this; if you had a lower proportion of photos that contained cats, you'd see this", and so on. By implication, it says: "You see this because you are a white male from Wisconsin who likes cats." That explanation may be true. It may even enable a user to develop an intuitive picture of how Facebook's News Feed ranking systems work (though we are sceptical even of that, since that intuitive picture applies only to their case and may not generalize).[30] But it is nonetheless beside the point. The individual wants to know why Facebook chose to construct its News Feed ranking system in the way it did. Why are engagement and relevance the primary metrics, and how are they defined? What are the other principles on the basis of which content is promoted and demoted on News Feed? They want Facebook to justify its News Feed ranking system. The kind of explanation this requires is on the level of choices and principles in the design of content moderation systems, not of interpretable machine learning models. Such technical explanations can actually distract from the appropriate form of justification. Citizens feel they no longer need to press for answers to the harder, but more fundamental question: Why do you distribute information in this way?

The posing of these questions by citizens, and the answering of them by institutions, is essential to the functioning of modern democracies. For large internet companies in particular, the drive towards transparency, cashed out in the form of the search for interpretable machine learning models, represents a welcome distraction from a fundamental debate about their own powers and purposes. The danger is that we make the same mistake in explanation and interpretability as we have in privacy: individual 'understanding' of a model takes the role 'consent' is supposed to play in securing important forms of institutional accountability. Individual understanding may often be just as much of an illusion as individual consent.[31]

---

[30] Along the lines of the kind of interactive explanation systems about which Edwards and Veale are more optimistic. *See generally* Edwards & Veale, *Slave to the Algorithm*, *supra* note 5.

[31] *See generally* Lilian Edwards, *Privacy, Law, Code and Social Networking Sites*, RESEARCH HANDBOOK ON GOVERNANCE OF THE INTERNET (2013); Rikke Frank Joergensen, *The Unbearable Lightness of User Consent*, 3 INTERNET POL'Y REV. (2014);

Accountability is constitutive of democratic self-governance. It is part of what it means for citizens to authorise in an ongoing way the complex decision-making procedures to which they are subject. Accountability requires that an institution justify its choices for the design and implementation of its decision-making procedures, including the use of statistical techniques and machine learning, to an individual or institution with meaningful powers of oversight and enforcement. The right form of explanation can be crucial to the giving and receiving of that justification. The wrong form can unintentionally or intentionally undermine it. Technical explanations of machine learning models can further the aim of institutional justification, and therefore of accountability. But they can also undermine and distract from it. The form of explanation should depend on the form of accountability. Institutional context should drive the form of explanation offered. We cannot simply adopt technical solutions to explanation without thinking through what is required for genuine accountability over time. It is to this challenge, and to the interpretation of the GDPR with this aim in mind, that we now turn.

II. SYSTEMIC ACCOUNTABILITY, JUSTIFICATION, AND THE GDPR

We now turn to the GDPR and the RtE debate that has dominated its reception amongst scholars in the U.S. The GDPR, which came into effect in May 2018, lays down requirements with respect to the information individuals must receive about automated decision-making in their case.[32] Several recent proposals have followed suit, seeking to ensure that machine learning models, which might otherwise be uninterpretable, can be explained to those whose lives they will

---

Elizabeth Denham, *Consent Is Not the 'Silver Bullet' for GDPR Compliance*, INFO. COMM'R OFF. NEWS BLOG (August 16, 2017), https://ico.org.uk/about-the-ico/news-and-events/blog-consent-is-not-the-silver-bullet-for-gdpr-compliance/.

[32] *See generally* Kaminski, *The Right to Explanation, Explained*, *supra* note 1; Casey et al., *Rethinking Explainable Machines*, *supra* note 1; Isak Mendoza & Lee Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, EU INTERNET L.: REG. AND ENFORCEMENT (2017); Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation,"* 38 AI MAG. 50, 50–57 (2017); Malgieri & Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, *supra* note 6; Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23; Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT. DATA PRIV. L. 76, 76–99 (2017); Sandra Wachter et al., *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841 (2018).

inevitably shape.[33] Broadly, the GDPR requires individuals to be provided with "meaningful information about the logic involved" in the automated decision[34] as part of the right to contest these decisions and to enforce other rights under the GDPR.[35]

There has been fierce disagreement about the scope and content of this explainability requirement. The core of the right to explanation in the GDPR regime can be found in Article 22 and Articles 13, 14, and 15. Article 22 lays down the general assumption against "automated individual decision-making, including profiling" and articulates the three exceptions to that assumption, while Article 13, Article 14 and Article 15 discuss the various transparency rights that arise from the use of automated decision-making, including the right to explanation. Article 13 creates requirements at the time information is collected from an individual, Article 14 focuses on requirements at the time information is collected from a third party, and Article 15 creates ongoing requirements related to the holding of individuals' information. These Articles bear on cases of decisions "based solely on automated processing" which "produce[s] legal effects concerning him or her or similarly significantly affects him or her," as they require the individual to be informed of the existence of the automated decision-making and for the provision of "meaningful information about the logic involved" in the automated decision.[36]

Our aim is not to offer another interpretation of this requirement. We agree with scholars who have recently argued that the GDPR's main text must be read alongside surrounding 'soft-law.'[37] These include the preamble to the Directive, known as the Recitals. These Recitals are not strictly binding, but they indicate how the GDPR is likely to be enforced and how, therefore, companies are likely to shape their behaviour to comply with the GDPR.[38] They also include the guidance of the Article

---

[33] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5; Edwards & Veale, *Slave to the Algorithm*, *supra* note 5.

[34] *See* GDPR arts. 13(2)(f), 14(2)(g) and 15(1)(g).

[35] Selbst and Powles make explicit this connection between providing information and individuals enforcing their rights. Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23.

[36] An individual also has the right to contest these decisions under Article 22. GDPR, *supra* note 21.

[37] *See generally* Kaminski, *Binary Governance*, *supra* note 1; Kaminski, *The Right to Explanation, Explained*, *supra* note 1; Casey et al., *Rethinking Explainable Machines*, *supra* note 1.

[38] As Kaminski puts it, "[t]hese texts are not technically binding, but they provide clarity of what is to come." Kaminski, *supra* note 1, at 195; In contrast, Wachter et al., who argue that the Recitals are not binding in the case of establishing the right to explanation since they are only

29 Working Party (A29WP), an advisory board made up of data protection authority representatives of all EU Member States, the European Data Protection Supervisor, and the European Commission. The purpose of the A29WP and its successor, the European Data Protection Board, is to promote consistent application of the GDPR across Member States.[39] Furthermore, the GDPR is designed to be given force by national Data Protection Authorities (DPAs), like many other EU Directives. How those institutions interpret its provisions is, in the end, what matters. We therefore give particular weight to guidance subsequently issued by national DPAs, most notably, the Information Commissioner's Office (ICO) in the U.K.[40]

In our view, this accompanying guidance makes it clear that the GDPR does contain a right to explanation. But more importantly, that guidance should shape how we elaborate on the content and scope of that right to explanation. The guidance suggests that the GDPR has begun to develop a comprehensive set of provisions for attaining systemic accountability over time. What a right to an explanation means in the context of the GDPR should depend on how the GDPR aims to secure systemic accountability.

Our aim is to approach the challenge of explainability by keeping in mind what is of ultimate importance: holding those with power to account, by ensuring that institutions justify their design and use of machine learning models to regulatory bodies and to individuals subject to their predictions, classifications, and rankings. The appropriate form of explanation should depend on who is justifying what to whom, as part of the process of accountability. To draw out the implications of this argument for interpreting the GDPR, we propose a simple taxonomy of justifications. It is broken down by three questions: (1) *Who* is offering

---

meant to provide guidance in cases of ambiguity, which is not the case, they argue, with respect to Article 22. Moreover, they argue that the Recital could not be used to establish new legal rights and duties that do not clearly exist in the text of the Directive. *See* Wachter, et al., *supra* note 32, at 80.

[39] GDPR replaced the pre-existing EU Directive on privacy, the Data Protection Directive, which came into force in 1995, and was suspended when the GDPR became enforceable in 2018.

[40] *See generally* ICO, *Guide to the General Data Protection Regulation (GDPR)*, U.K. INFO. COMM'R OFF. (August 2018), https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf; *See also* ICO, *Rights Related to Automated Decision Making Including Profiling*, U.K. INFO. COMM'R OFF. (2017), https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/; *see also* ICO, *Big Data, Artificial Intelligence, Machine Learning and Data Protection*, U.K. INFO. COMM'R OFF. (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

the justification?; (2) *What* is being justified?; and (3) *To whom* is the justification offered?

Who should offer the justification?

We leave this question to one side. It implies distinctions between both between engineers who design the model and the institutions that use it as part of their decision-making process, and between different forms of institutions, such as private and public. We focus on the justification of decision-making processes by private institutions.

What should be justified?

1. The machine learning model (overall logic or specific predictions); OR

2. The choices an institution makes about the design of a machine learning model and its integration into their decision-making procedure.

To whom should the justification be offered?

A. An individual subject to the model's predictions, classifications or rankings; OR

B. A regulator or some other type of public oversight body.

We focus on two categories of justification that can be drawn from this taxonomy. The first is 1A. The explanation of a machine learning model (1) to an individual (A). The second is 2B. The explanation of the choices an institution makes in the design and implementation of a decision-making procedure (2) to a regulator or some other public oversight body (B). Let's take each in turn.

The debate about whether the GDPR contains a RtE focuses on the 1A category. It concerns whether an individual has a right to "meaningful information about the logic involved" in a fully automated decision which "significantly affects him or her."[41] This has produced a range of approaches to explaining machine learning models to individuals that would satisfy this requirement, from straightforward

---

[41] Selbst & Powles, *Meaningful Information and the Right to Explanation, supra* note 23.

counter-factual explanations[42] to more complex technical approaches to developing interpretable models. These technical approaches aim to summarise the logic of a complex machine learning model in a simpler, more comprehensible model. Most explain how machine learning models work after the fact, known as reverse engineering. These tend to either summarise the whole logic of the model, known as global approaches, or to explain a specific set of outcomes the model produces, known as local approaches.[43]

We believe this focus on the 1A category is mistaken. The 1A category, the requirement that an institution explain how its machine learning model works to an individual subject to those decisions, is not a satisfactory way of holding institutions to account. Knowing what the rules are is not itself a check on the power of those who decide what the rules are. The category mistakenly characterises a challenge of institutional justification as a challenge of algorithmic explanation. Focusing on the requirement of those with power to inform subjects as to what the rules are, intentionally or not, distracts from the higher-order question of what the rules should be. If Facebook offers a tool that allows an individual to understand why their News Feed shows them what it does, the danger is that the user feels as though Facebook has justified its more general choices about how it distributes information on News Feed. It has in fact done nothing of the sort. It suits Facebook for the debate to focus on how they can develop technical explanations of News Feed's ranking models, rather than on the principles Facebook chooses to impose on its content moderation systems. The latter draws attention to Facebook's underlying power to decide who sees what, and why.

Nor is the 1A category a satisfactory interpretation of the GDPR's most important provisions. The GDPR contains important mechanisms for systemic accountability, which focus on forcing an institution to

---

[42] *See* Wachter et al., *supra* note 32, at 854.

[43] *See generally* Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS 1, 1–42 (2018); Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE AND INFO. SYS. 95, 95–122 (2018); Selbst and Barocas *supra* note 3; Zachary Lipton, *The Mythos of Model Interpretability*, CORNELL UNIV. (2017), https://arxiv.org/abs/1606.03490; Kroll et al., *supra* note 33; Jatinder Singh et al., *Responsibility & Machine Learning: Part of a Process* (Working Paper, 2016), https://papers.ssrn.com/abstract=2860048; Marco T. Ribeiro et al., *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, 22 ACM 1135, 1135–1144 (2016); Tameru Hailesilassie, *Rule Extraction Algorithm for Deep Neural Networks: A Review*, 14 INT'L J. COMP. SCI. & INFO. SEC. 376, 376–380 (2016); Anupam Datta et al. *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, 37 IEEE SYMP. ON SEC. & PRIV. 598, 598–617 (2016).

justify their choices in the design and implementation of algorithmic decision-making systems, including their broader policy and commercial aims. Read in conjunction with the accompanying guidance of the Recitals, and the guidance published by A29WP and the ICO, the GDPR contains provisions that have the potential to transform the *ex-ante* process of designing machine learning models and integrating them into the decision-making systems of a range of important institutions. It sets out clear mechanisms for structuring systemic accountability, to ensure institutions justify the choices they make in that process. These include empowered DPAs, broad Data Protection Impact Assessments (DPIAs), auditing, and ethical review boards.[44]

This section uses our taxonomy of justifications to explore what this broader, more expansive reading of the GDPR implies for various forms of explanation. We contrast our 1A and 2B categories—the explanation of a machine learning model to an individual and the explanation of the decisions made in the design and implementation of that model to a regulator—to explore what is wrong with the more limited readings of the GDPR's provisions. The aim is to learn some broader lessons about the governance of institutions designing decision-making systems that use machine learning.

A. *What Should Be Justified: Institutions and the Process of Machine Learning*

We first focus on what it is that should be justified in the process of securing systemic accountability in the governance of algorithmic decision-making. If it is the machine learning model itself that must be justified, it would seem to follow that such a justification depends on an explanation of how the model works, either in terms of its overall logic or some subset of specific predictions.

This reasoning is mistaken, but it is encouraged by the text of the GDPR itself. Article 22 focuses on decisions "based solely" on automated data processing. The question of what exactly this means has divided scholars. Some have argued that decision-making procedures which involve humans in some perfunctory way would be exempt from Article

---

[44] *See* Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 208 ("Accompanied by other company duties in the GDPR—including establishing data protection officers, using data protection impact assessments, and following the principles of data protection by design—this regime, if enforced, has the potential to be a sea change in how algorithmic decision-making is regulated in the EU.").

22's requirements.[45] Much more persuasively, others argue that human involvement must be meaningful, as the A29WP guidance states, involving a person who has the "authority and competence to change the decision."[46] Article 22 in fact creates a strong presumption, or even prohibition, against solely automated decision-making, subject to three exceptions.[47] The GDPR intends to target decision-making systems that are fully automated, those which are, for instance, wholly constituted by a machine learning model. The right to explanation applies to these cases only.[48]

Articles 13, 14, and 15 then require that the data controller provide information about "the logic involved" in the automated decision-making. Here again, the language of the text itself is ambiguous. It is likely that this involves a requirement to explain the logic of the whole machine learning model rather than a subset of the predictions it produces.[49] If so, the GDPR is broader than other legal requirements to explain automated decisions, such as the requirement in the Equal Credit Opportunity Act (ECOA) that an applicant be provided a "statement of specific reasons for the action taken."[50] The ECOA requirement focuses on the individual outcome only, while the GDPR arguably requires a broader form of explanation.

This would seem to produce a view of the resulting right to explanation that falls squarely within the 1A category. The GDPR, on this view, requires an explanation of the logic of an entire machine learning model,

---

[45] *See* Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, *supra* note 32, at 88 ("Quite crucially, this creates a loophole whereby even nominal involvement of a human in the decision-making process allows for an otherwise automated mechanism to avoid invoking elements of the right of access.").

[46] *See* Casey et al., *Rethinking Explainable Machines*, *supra* note 1, at 171 ("According to the A29WP, companies must ensure that any human 'oversight of [a] decision is meaningful, rather than just a token gesture' if they intend for their systems to fall outside the scope of Article 22's provisions pertaining to decisions 'based solely on automated processing.'").

[47] These exceptions are: consent, contract, or if authorised by Union or member state law. *See* Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 197-198 (describing the three exceptions to the Article 22 right and prohibition); Isak Mendoza and Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling* 14 (U. OSLO FAC. L. LEGAL STUDIES, Research Paper No. 2017-20, 2017) (providing the exceptions to the Article 22 right).

[48] "Interpreting Article 22 as a prohibition rather than a right to be invoked means that individuals are automatically protected from the potential effects this type of processing may have." Article 29 *Data Protection Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation* 2016/679 (Feb. 6, 2018), at 20 [hereinafter A29WP]. Also note that, according to the Guidelines, the exceptions in Article 22 should be interpreted narrowly. *Id.* at 13.

[49] *See* Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23, at 236.

[50] 12 C.F.R. § 1002.9(a)(2)(i).

where that model constitutes the whole decision-making procedure that results in legal or similarly significant effects on a data subject.

This is not only a limited reading of the provisions and intent of the GDPR, it also completely misunderstands the role that explanation should play in a broader system for structuring accountability in the governance of algorithmic decision-making. Machine learning is a way of establishing a decision-making procedure. It is best thought of as a process, one that involves choices at every stage. These choices are made by institutions who design and integrate machine learning models into their decision-making procedures. These choices profoundly shape the form the machine learning model takes, the role it plays in their decision-making procedures, and the effects those decisions have on individuals over time. We believe that the RtE should be read in the context of the GDPR's broader provisions for mechanisms to secure accountability over time. These focus on the ex-ante design and implementation of decision-making procedures using machine learning.[51]

There are three crucial choices in the process of machine learning itself that must be considered, along with a broader set of choices about the role the machine learning model plays in the decision-making procedure, and the policy or commercial aims the institution has in deploying it.

1. Outcome of Interest

First, the outcome of interest is what the machine learning model looks for, that is, what it predicts, ranks, or classifies. The selection of an outcome of interest very often embeds important moral and political choices, which profoundly shape the predictions, classifications, or rankings the model will produce.[52] This choice, and the reasons for making it, require justification.

2. Training Data

Second, the training data set is what the machine learning models from. Recent research has developed several technical approaches to the evaluation of fairness in training data.[53] There are multiple aspects to the selection and construction of a training dataset, all of which can be

---

[51] *See infra* note 52.

[52] *See generally* Cary Coglianese and David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L. J. 1147 (2017).

[53] *See generally* Rich Zemel et al., *Learning Fair Representations*, 2013 INT'L CONF. MACH. LEARNING 325 (2013); J. Henry Hinnefeld et al., *Evaluating Fairness Metrics in the Presence of Dataset Bias*, CORNELL UNIV. (Sept. 24, 2018), http://arxiv.org/abs/1809.09245.

extremely important in shaping the predictions of the resulting machine learning model. These range from choices about time periods, demographic representativeness, and how to label the data.

3. Features

Third, the features included in a machine learning model. This includes choices about whether to include or exclude protected traits, such as race and gender. Removing a protected trait from a model is neither necessary nor sufficient to prevent discrimination in machine learning. In fact, preventing discrimination may require that information about individual membership of protected groups be *included* in machine learning models; fairness might require awareness, not blindness.[54] It also includes choices about whether to simplify the model by reducing the number of variables.[55]

Accountability requires justification. The form of explanation that justification requires depends on who is justifying what to whom. The GDPR is concerned with holding to account institutions which use automated decision-making procedures in important spheres.[56] Technical explanations of the logic of a machine learning model to an isolated individual will not be conducive to the kind of ongoing accountability the GDPR requires. The very form a machine learning model takes depends on choices made by humans in its design and implementation. The notion of providing a technical explanation of a machine learning model completely obscures the important and prior question: How did the rules that govern the operation of the automated decision come to be what they are? That is a question about the justification of institutional choices which is both prior to and much more significant than the question of what the

---

[54] *See generally* Cynthia Dwork et al., *Fairness through Awareness*, 2012 PROC. INNOVATIONS. IN THEORETICAL COMPUT. SCI. 214 (2012); Talia B. Gillis and Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459, 471 (2019); Symposium, Nina Grgic - Hlaca et al., *The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making*, 29 CONF. ON NEURAL INFO. PROCESSING SYS. (2016).

[55] Veale et al. describe a case in which the performance target of 75 percent was specified in advance, so the number of features could be reduced from 18,000 to 200, then 20, then 8, "because it's important to see how it works, we believe." Michael Veale, Max Van Kleek, and Reuben Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making* 440, PROC. OF THE 2018 CHI. CONF. ON HUM. FACTORS IN COMP. SYS. (2018), http://arxiv.org/abs/1802.01029.

[56] For instance, the 'spheres' in which DPIAs might be required are described as 'high-risk' in the text. The A29WP guidance lists a set of concrete criteria that make clear the broad scope of what 'high-risk' might mean. *See* Casey et al., *supra* note 1, at 176 ("Article 35(7) of the GDPR enumerates four basic features that all DPIAs must, at a minimum, contain."); A29WP, *supra* note 48.

rules are. It is also, we have argued, a question to which the GDPR's provisions aim to elicit an answer.

This is precisely what the A29WP guidance states. The guidance explains that "the complexity of machine-learning" algorithms "can make it challenging to understand how an automated decision-making process or profiling works."[57] Such complexity, the guidance continues, "is no excuse for failing to provide information."[58] Companies whose decisions are subject to the provisions of Article 22 "should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision," "not necessarily a complex explanation of the algorithms used or [a] disclosure of the full algorithm."[59] The guidance further clarifies that this will include information used in the decision-making process, including: categories of data; the source of that information; how many profiles were constructed and used in the procedure; and how that profile is used for a decision about the data subject.[60]

Institutions always make choices about how to design and integrate machine learning models into their decision-making procedures. In these choices lie trade-offs about discrimination and fairness, who wins and who loses, along with a host of other normative and epistemological assumptions. It is for these choices that an institution must be held accountable. The GDPR's provisions for a RtE must be understood in this context. Surrounding guidance makes clear that the appropriate form of explanation is not specifically about the logic of the machine learning model, but the choices an institution made in designing and integrating it into their decision-making system.[61]

B. *To Whom Should the Justification Be Offered: Regulators and Citizens*

There is also confusion about to whom the justification is owed. Here again, the language of the GDPR is not helpful. The GDPR text itself does not explain the aims of a RtE. However, the guidelines explain that "the data subject will only be able to challenge a decision or express their view

---

[57] *Id.* at 25.

[58] *Id.*

[59] *Id.*

[60] *Id.* at 31.

[61] For a useful overview of the kinds of choices that might be required for the form of justification at which the GDPR aims, which they term 'legibility,' *See generally* Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, *supra* note 6.

if they fully understand how it has been made and on what basis."[62] The emphasis on the ability to challenge the decision reflects the fact that on this view, the purpose of the explanation is to invoke a data subject's other fundamental rights. As Kaminski puts it, "[i]ndividual transparency provisions, as the guidelines make clear, are intended to empower individuals to invoke their rights under the GDPR."[63]

We think this is a problem not with scholarly interpretations of the GDPR, but with the reasoning of the text itself and the guidelines surrounding it. The idea that the disclosure of information produces the enforcement of rights is not supported by evidence. Other areas of consumer behaviour research suggest people often struggle understanding straightforward information about products and how they pertain to their personal information.[64] The GDPR's instrumental individual transparency approach goes one step further, assuming that individuals will not only understand the information they are provided, but also that they will recognize violations of their legal rights and act on them.[65] Furthermore, many of the fundamental concerns about using machine learning to make decisions – most notably those related to bias and discrimination – can only be understood with a systematic and aggregate analysis of the decision-making procedure. The explanation of an individual decision to an isolated individual will not enable this kind of aggregate analysis; in fact, it may even obscure demands for obtaining it. The GDPR's account of the instrumental aim of an individual RtE is not convincing.

If systemic accountability is placed front and centre, rather than individual rights, it is clear that institutional justification of decision-making procedures must be offered to empowered, well-resourced regulators. There are ample provisions in the GDPR for doing just this. The individual RtE should not distract or detract from these provisions for systemic accountability. Rather, as we have consistently argued, the RtE should be viewed as a means to this broader end.

A focus on systemic accountability produces a very different view of the kind of explanations a regulator might require from an institution. We believe that at minimum, an explanation that supports the form of justification required by systemic accountability would answer the following questions. In all cases, the institution must not only provide a satisfactory answer to the question, it must provide reasons for the answers

---

[62] A29WP*, supra* note 48, at 27.

[63] Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 211.

[64] *See supra* note 24. *See also* Oren Bar-Gill and Kevin Davis, *(Mis)perceptions of Law in Consumer Markets*, AM. L. & ECON. REV. (2017) (discussing misperceptions of the law, which is an additional reason that disclosures alone may be insufficient).

[65] *See* Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5, at 52.

given. Where relevant, answers could be accompanied by quantitative data and analysis.

1. *What are the goals of the decision-making procedure?*

2. *What are the company policies that constrain or inform the decision-making procedure, including the role machine learning plays within it?*

3. *How did the company define the outcome of interest the machine learning model was trained to predict? Why?*

4. *How did the company select and construct the data on which the model was trained? If relevant, how was the data labelled and by whom? Was the impact of using other training data considered?*

5. *What features did the company choose to include or excluded in the model? Why?*

6. *Does the decision-making procedure involve human discretion? How precisely do the automated and human element of the decision-making procedure interact? Has the company considered how this interaction effects aggregate outcomes?*

7. *Has the lender considered how this interaction affects decisions?*[66]

The GDPR has ample mechanisms for encouraging, if not requiring, companies to answer these questions. As Kaminski argues, rather than "arguing over" the "instrumental value of individual notice, or publicly releasing source code," we should be debating how to obtain structured "accountability across a firm's decision-making, over time."[67]

Consider Data Protection Impact Assessments (DPIAs).[68] DPIA's are a "process for building and demonstrating" compliance by systematically

---

[66] For an alternative and insightful list of questions, s*ee generally* Malgieri and Comandé, *supra* note 6, at 29-30.

[67] Kaminski, *Binary Governance, supra* note 1, at 35.

[68] There are others mechanisms in the GDPR for attaining systemic accountability, such as auditing and ethical review boards. *See e.g.* Kaminski, *Binary Governance, supra* note 1, at 8 ("The instrumental rationale for regulating algorithmic decision-making counsels that regulation should try to correct these problems, often by using systemic accountability mechanisms, such as ex ante technical requirements, audits, or oversight boards, to do so."); Kroll et al., *supra* note

examining how automated decision-making procedures are designed and implemented. They are meant to be an "iterative process" that fall within the GDPR's broader "data protection by design" principles, which apply throughout the design, implementation and monitoring of a decision-making procedure.[69] DPIAs are more than simple recommendations of best practice. They are intended to apply to a broad range of institutions which use data to make important decisions. Importantly, those decisions must not be solely automated. As the A29WP guidance states, DPIAs apply "in the case of decision-making including profiling with legal or similarly significant effects that is not wholly automated, as well as solely automated decision-making defined in Article 22(1)."[70] Where appropriate, companies should "seek the views of data subjects or their representatives" during the DPIA process.[71] And companies should explain their reasons for making the choices they did in the design and implementation of their models.

In this context, the scope and content of the RtE is much broader. As Casey et al. argue, the right to explanation "is no mere remedial mechanism to be invoked by data subjects on an individual basis, but implies a more general form of oversight with broad implications for the design, prototyping, field testing, and deployment of data processing systems."[72] We agree with Veale and Edwards that *ex ante* DPIAs will "become the required norm for algorithmic systems, especially where sensitive personal data, such as race or political opinion is processed on a large scale."[73]

This is as it should be. The form of explanation required for institutional justification will often not be the technical explanation of the logic of machine learning models to isolated individuals. This is the 1A

33, at 660 ("Beyond transparency, auditing is another strategy for verifying how a computer system works."); Selbst & Barocas, *supra* note 3, at 1133 ("The most common trigger of the latter is a lawsuit, in which documents can be obtained and scrutinized and witnesses can be deposed or examined on the stand, but auditing requirements are another possibility.").

[69] *See* Casey et al. *supra* note 1, at 172-173; A29WP, *supra* note 48, at 29 ("As a key accountability tool, a DPIA enables the controller to assess the risks involved in automated decision-making, including profiling. It is a way of showing that suitable measures have been put in place to address those risks and demonstrate compliance with the GDPR.").

[70] *Id.* at 32 ("The following list, though not exhaustive, provides some good practice suggestions for controllers to consider when making solely automated decisions . . . ."). *See also* Casey et al., *supra* note 1, at 174 (According to the Regulation, DPIAs are mandatory '[w]here a type of processing[,] taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons.'") (internal citations omitted).

[71] *Id.* at 36.

[72] *Id.* at 39.

[73] Edwards & Veale, *Slave to the Algorithm?, supra* note 5, at 78.

category. Rather, it should be an explanation of the decisions an institution made in the design of a machine learning model and its integration into their decision-making procedure, to an empowered regulator. This is the 2B category. Reporting to a regulator rather than to an individual is necessary to reveal aggregate patterns and effects that are not discoverable when considering a decision in isolation.[74] Regulators and other public bodies have the technical knowledge, skills and time to evaluate information that an individual does not.[75] The very purpose of regulators is to take actions in situations when it is individually not worthwhile, but is socially desirable.

CONCLUSION

The RtE debate should begin with the foundational goal: accountability. Accountability is constitutive of democratic self-governance. It is an integral aspect of a citizenry's ongoing authorization of the complex decision-making systems which shape their lives. Part of what it means to be a citizen of a self-governing polity is to give and receive justifications of those decision-making systems. Explanations are

---

[74] One of us has written about this type of aggregate analysis elsewhere when considering the type of information a lender would provide to the CFPB to allow testing of whether credit pricing algorithms are compliant with discrimination law. *See generally* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019). In the context of credit pricing discrimination, this has been one of the most significant barriers to a successful discrimination complaint. The passing of the Home Mortgage Disclosure Act in 1975, increased the ability to bring a successful discrimination claim and class action against lender since the Act mandated the disclosure of mortgage applications and their outcomes, allowing for an aggregate consideration of mortgage decisions. *See e.g.* Robert G. Schwemm & Calvin Bradford, *Proving Disparate Impact in Fair Housing Cases after Inclusive Communities*, 19 N.Y.U. J. LEGIS. & PUB. POL'Y 685, 713-15 (2016).

[75] Future technical research into explainability and interpretability in machine learning could benefit from assuming that the appropriate audience for their approaches is not isolated individuals but regulators. The great strength of Dwork's 'individual fairness' approach is that it isolates the normative choices and therefore makes possible a form of accountability, e.g. fair affirmative action, through the choice of the distance metric. It can require access to protected status information during the design phase, usually explicitly prohibited, which may require a big shift in policy. What matters though is a *procedure* which justifies the choice of the distance metric, which can be explained to either a regulator or, in some cases, those who are actually subject to the decision. *See* Dwork et al., *supra* note 54, at 2 (describing the "[c]onnection between individual fairness and group fairness," Dwork et al. state that "[s]tatistical parity is the property that the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole. Statistical parity speaks to group fairness rather than individual fairness, and appears desirable, as it equalizes outcomes across protected and non-protected groups."); *see also id.* at 3 ("Justifying the *availability* of or access to the distance metric in various settings is one of the most challenging aspects of our framework, and in reality the metric used will most likely only be society's current best approximation to the truth.").

valuable insofar as they are required to achieve systemic accountability over time. In practice, this means that the appropriate form of explanation will depend on who is justifying what to whom. We have argued that the RtE debate focuses far too much on the explanation of the logic of a machine learning model to isolated individuals. What matters for accountability is the justification by an institution of the choices it made in the design and implementation of a machine learning model. The form of systemic accountability should drive the form of institutional justification, which in turn, should drive the appropriate form of explanation.

Interpreting the GDPR matters because it is likely to shape future regulation of algorithmic decision-making. The primary concerns that arise when using machine learning to make, or assist with, important decisions are not satisfactorily addressed by focusing on the rights of isolated individuals, or the logic of an individual machine learning model itself. As we develop comprehensive governance structures to address the concerns that arise from the use of machine learning in decision-making, we should move beyond frameworks that rely on the individual enforcement of rights, and towards those which develop a systemic approach to establishing and maintaining accountability within a complex modern democracy.

This means moving beyond privacy as a lens through which to view the governance of algorithmic decision-making. Some of the limited ways in which the GDPR has been interpreted have been transplanted from older debates about privacy. This is partly because the GDPR itself grew out of earlier privacy provisions and it is partly because scholars who interpret it often cut their cloth in the privacy field. The focus on individual rights, as well as the notice and consent framework that underpins the GDPR's approach, are all characteristic of approaches to addressing concerns about privacy. As Kaminski puts it, "the strong system of individual rights" within the GDPR may come "at the cost of correcting systemic problems essential for achieving accountability in modern democracies."[76] If the RtE is interpreted as requiring explanations of the logic of machine learning models to isolated individuals, these explanations are not likely to be useful to regulators in evaluating whether to accept the justification of an institution of its decision-making procedure. That is, such

---

[76] Kaminski, *Binary Governance*, *supra* note 1, at 74. This also means relating current discussions about the governance of algorithmic decision-making to a rich literature on regulatory strategies in an administrative state. *See e.g. id*. at 30-31 ("If there is already concern in administrative law over insulating government bureaucrats from electoral and judicial oversight, collaborative governance compounds such concerns by involving private parties.")

explanations may actually obstruct systemic accountability. Most challenging of all, the GDPR requires companies to assist in the enforcement of citizens' fundamental rights. This effectively privatizes the protection of individual rights. The GDPR and the literature surrounding it has no satisfactory account of how its provisions are to be subject to democratic oversight. Accountability matters because it is constitutive of collective self-government. Future regulatory provisions must focus more directly on developing mechanisms within modern democracies that can secure accountability in the governance of algorithmic decision-making systems.

We are currently in a moment of choice. We are choosing how to integrate humanity's most powerful decision-making tool – machine learning – into a range of complex human activities. We have argued that institutional justification, not algorithmic explanation, is essential to the accountability constitutive of democratic self-government. The technical explanation of machine learning models is never sufficient, is often not necessary, and sometimes actively distracts from, the justification of the decision-making systems of which they are a part. We must think through what it means to reason about the justifications an institution should offer for its choices in how and why it constructed its decision-making procedure in the way it did – that is, a justification of why the rules are what they are. We have offered a sketch of what such a system of reasoning might look like.

We must keep our eyes on the right prize. That prize is accountability. Institutional power is held in check by other institutions with the authority and resources sufficient to hold them to account. To attain that prize requires a laser-like focus on choice in the face of apparent technical inevitability. In this case, it means requiring institutions to justify their choices about how they have constructed their decision-making systems. Not being distracted by whizzy technical explanations of their machine learning models work – or even, of that most dangerous of terms, artificial intelligence.