AIとプライバシーに関する法的課題の最近の動向

2023年3月16日
@産総研・人工知能セミナー

水野祐（弁護士、シティライツ法律事務所）

自己紹介

- **法律家／弁護士**（シティライツ法律事務所）

- Creative Commons Japan理事
- Arts and Law理事
- 九州大学GIC客員教授／慶應義塾大学SFC非常勤講師
- グッドデザイン賞審査員
- note株式会社などの社外役員
- 著作
  - 『法のデザイン ―創造性とイノベーションは法によって加速する』（フィルムアート）
  - 『オープンデザイン参加と共創から生まれる「つくりかたの未来」』（オライイリー・ジャパン、共同翻訳・執筆）
  - 『新しい社会契約（あるいはそれに代わる何か）』（WIRED連載）など

https://twitter.com/TasukuMizuno

シティライツ法律事務所 CITY LIGHTS LAW

MISSION

ABOUT
MISSION
TEAM
NEWS

法を駆使して
創造性、イノベーションを最大化する

シティライツ法律事務所は、After the Internetのテクノロジー、ビジネス、カルチャーに対する深い理解を前提としたリーガルサービスを提供する法律事務所です。私たちは、法律や契約といった法に関する様々なスキルや経験を駆使して、クライアントの創造性やイノベーションを最大化することをミッションとして掲げています。法律家としての専門性を当然の前提としつつも、既存の枠組みや固定観念にとらわれない「姿勢」により選択される、個性を持った少数精鋭の法律事務所を志向しています。

弊所のメンバーは、インターネットを含むICT、メ

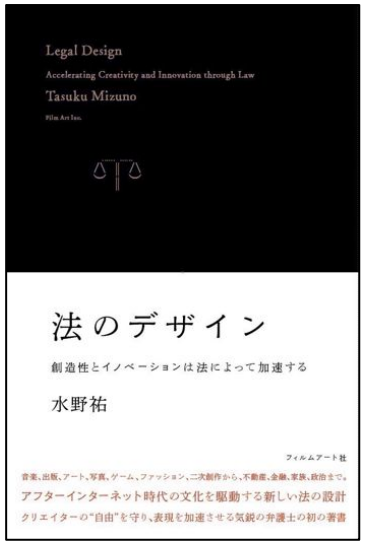
Using the legal framework to
maximize creativity and innovation

City Lights Law provides legal services with a deep, foundational knowledge of the technology and business of, and culture in, the post-Internet society. With a mission to maximize our clients' creativity and innovation, we aim to utilize our legal skills and capitalize on our experiences. While providing world-class legal expertise, we operate from a unique problem-solving stance that moves beyond pre-existing assumptions or rigidly fixed concepts.

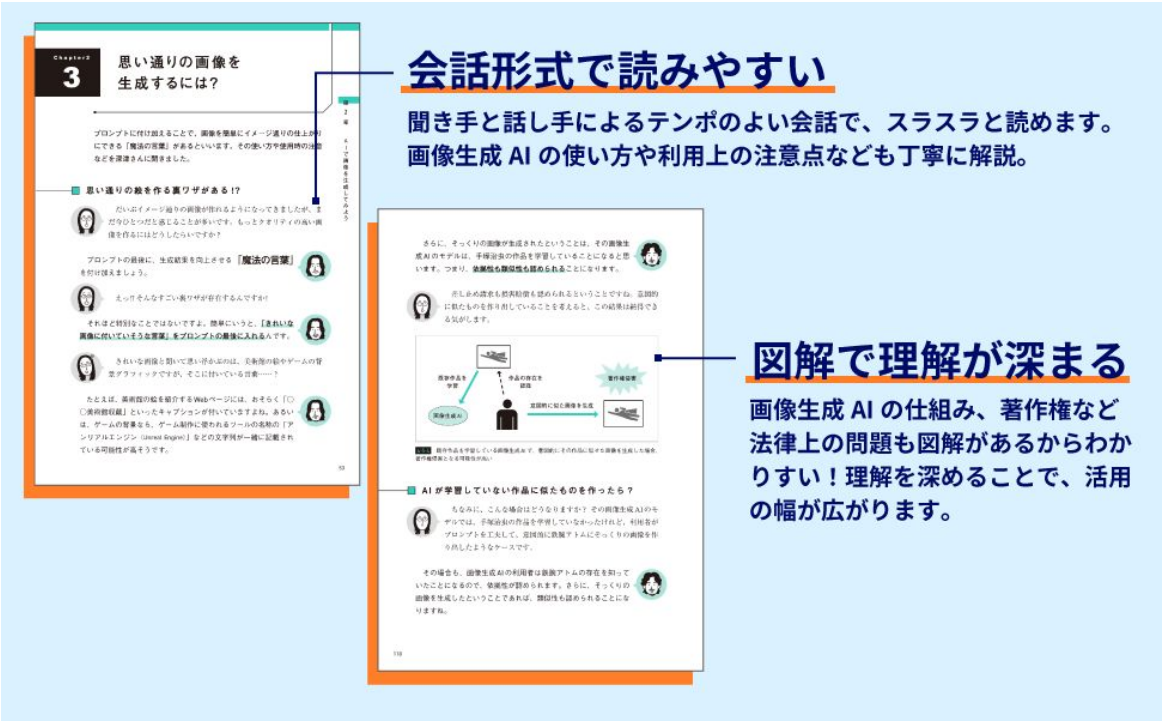
http://citylights.law/mission/

『法のデザイン 創造性とイノベーションは法によって加速する』（2017）

- 法（法律や契約など）を単に規制として捉えるのではなく、**物事や社会を良い方向に誘導したり、加速させたりするための滑走路・潤滑油・補助線**としても捉えるアプローチ
- 法とは、わたしたちの生活を「心地良く」「豊か」にするためのツールであり、テクノロジーであるとする立場
  - 「守る」ものではなく、「使う」もの
  - 見直して、アップデートするもの
- 「**リーガルデザイン**」という考え方
  - 立法や法改正だけでなく、法解釈（と行政との対話）や共同規制などの公民による法の共創から、契約、知財戦略も含む。
  - 既存のルールを疑い、新しいルールを設計・共創していく視点、姿勢、技術
- 時代とともにルールは変化していくことを前提として、一般市民・企業がルール形成に積極的に参加していく**ボトムアップ型のルール形成**システムを提案



深津貴之、水野祐、酒井麻里子『先読み！IT x ビジネス講座 画像生成AI』



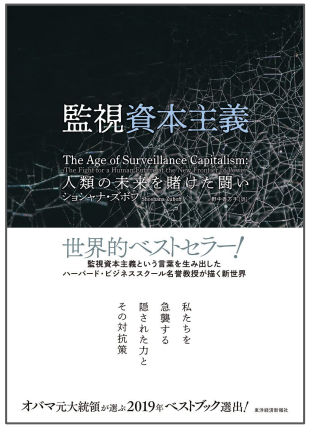
https://book.impress.co.jp/books/1122101128

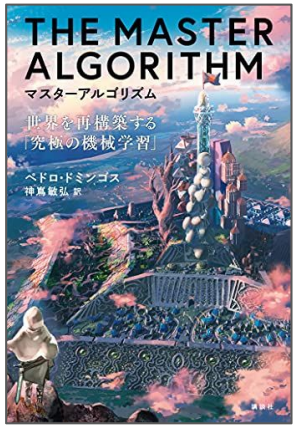
連載『新しい社会契約』（WIRED JAPAN）

https://wired.jp/magazine/vol_48/

AIとプライバシーを巡る現状認識と視座

- **現状認識**
  - **監視資本主義**（byショシャナ・ズボフ）
    - ビッグテックに代表されるごく一部の大企業のサービス設定により、利用者のプライバシー条件が左右される。
  - **新らしいネットワーク効果**（byペドロ・ドミンゴス）
    - データ寡占とAIのアルゴリズムにより「勝者総取り」の新しいネットワーク効果が生まれる
  - **プライバシーを巡るパラドクス（プライバシー・パラドクス？）**
    - パーソナルデータを積極活用してAIの予測精度を高めようとすればするほど、プライバシー侵害のリスクは高まる。逆も然り。
    - 「プライバシーを重視している」と言っているユーザー・消費者であっても、実際にはほとんど見返りなしにパーソナルデータを提供したり、プライバシー保護のための手段を講じない。
- **視座**
  - **権利とアーキテクチャによるプライバシーの実質的な保護**
    - 認知的・AI介入的な自己決定の限界
  - **競争法またはコーポレート・ガバナンスの視点**
    - パーソナルデータを提供している企業（特にプラットフォーム）とユーザーとの関係を受託信認関係と捉える視点（コーポレート・ガバナンスとしての視点）
  - **「思想の自由市場」破壊と民主主義**
    - フィルターバブル、エコーチェンバー

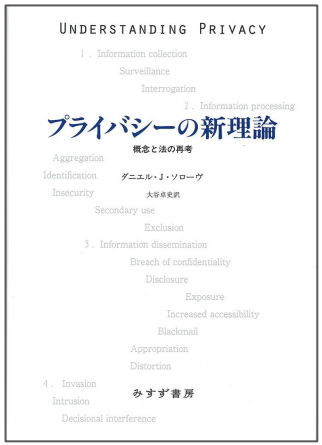




ショシャナ・ズボフ『監査資本主義』（東洋経済新報社、2021）
ペドロ・ドミンゴス『マスターアルゴリズム』（講談社、2021）

プライバシーとは

- プライバシーは時代、場所、文化等によって異なりうる**相対的な概念**
- ただし、**プライバシー（権）の核心をどのように捉えるかが、実務や法制度の動向に影響を与えうる**
  - プライバシー1.0（古典的プライバシー）：私生活上の秘密を不特定多数の第三者に公表・暴露されない権利（消極的な権利）
    - 日本では、「私生活をみだりに公開されないという法的保障ないし権利」（「宴のあと」事件）
  - プライバシー2.0（情報プライバシー）：自己情報コントロール権、情報自己決定権
    - 誰とどこまでの情報をシェア・共有するかを自ら主体的に選択・決定できるという情報共有の範囲に関する決定権が情報社会で自律的に生きていくために必要
  - **プライバシー3.0（構造的・関係的プライバシー）：アーキテクチャ志向型の自己情報コントロール権・情報自己決定権**
    - ネットワーク社会（ネットワークへの常時接続）を背景に、プライバシーを制限または実質的に保護するアーキテクチャの在り方が焦点に（ex. プライバシー・バイ・デザイン）
    - データポータビリティ権
    - 信頼または信認義務に基づくプライバシー保護論
  - プライバシー4.0？（パーソナルデータの客観的な取扱態様に着目し、自己決定の要素を積極的に抜く方向性の模索）：自己情報の適正な取扱いを受ける権利



ダニエル・J・ソロブ『プライバシーの新理論』（みすず書房、2013）

駒村圭吾編著『Liberty 2.0 自由論のバージョン・アップはありうるのか？』（弘文堂、2023）

AI開発・利用におけるプライバシー保護の実務的な留意点

- パーソナルデータの利用にあたっては、**個人情報保護法による行為規制への対応**と**プライバシー権・肖像権侵害への配慮**の両方の側面から検討することが求められる
- インターネットから収集（クローリング、スクレーピング）してきた学習用データ（セット）には、膨大なパーソナルデータが含まれている！
- 学習用データの取得の段階
  - 適正取得・利用義務
  - 「要配慮個人情報」（人種、信条、病歴など不当な差別または偏見が生じる可能性のある個人情報）は本人同意なく取得しない
- 学習済みモデルの段階で、**特定の個人との対応関係が排斥されている限りは個人情報に該当せず**、統計情報と同視できると整理すれば、個人情報保護法の規制にかからない。
  - **特定の個人との対応関係をきちんと排斥されているか否かはあやしいのでは？**
- 個人情報（個人データ）と整理せざるを得ない場合、本人同意の取得はほぼ不可能…
  - 学習用データセットの提供時に第三者の権利侵害がないことを保証してもらう契約を締結する？
  - なかなか締結してもらえないし、仮に締結してもらっても実効性に乏しい。
    - 業界の多くのプレイヤーが利用している学習用データだから大丈夫だろう？
    - 100万分の1程度の確率なのでリスクが低い？
- **利用規約・プライバシーポリシーにおいて、プライバシー侵害など違法な利用行為、学習データの復元行為などを禁止事項として定めておく必要**
- 個人情報保護法違反・プライバシー権侵害にならないとしても、**倫理的に問題ないかの確認も別途必要**

Extracting Training Data from Diffusion Models

Nicholas Carlini[*1] Jamie Hayes[*2] Milad Nasr[*1]
Matthew Jagielski[*1] Vikash Sehwag[*4] Florian Tramèr[*3]
Borja Balle[*2] Daphne Ippolito[*1] Eric Wallace[*5]
[1]Google [2]DeepMind [3]ETHZ [4]Princeton [5]UC Berkeley
*Equal contribution *Equal contribution *Equal contribution

Abstract

Image diffusion models such as DALL-E 2, Imagen, and Stable Diffusion have attracted significant attention due to their ability to generate high-quality synthetic images. In this work, we show that diffusion models memorize individual images from their training data and emit them at generation time. With a generate-and-filter pipeline, we extract over a thousand training examples from state-of-the-art models, ranging from photographs of individual people to trademarked company logos. We also train hundreds of diffusion models in various settings to analyze how different modeling and data decisions affect privacy. Overall, our results show that diffusion models are much less private than prior generative models such as GANs, and that mitigating these vulnerabilities may require new advances in privacy-preserving training.

1 Introduction

Denoising diffusion models are an emerging class of generative neural networks that produce images from a training distribution via an iterative denoising process [64, 66, 33]. Compared to prior approaches such as GANs [30] or VAEs [46], diffusion models produce higher-quality samples [18] and are easier to scale [56] and control [51]. Consequently, they have rapidly become the de-facto method for generating high-resolution images, and large-scale models such as DALL·E 2 [56] have attracted significant public interest.

The appeal of generative diffusion models is rooted in their ability to synthesize novel images that are ostensibly unlike anything in the training set. Indeed, past large-scale training efforts "do not find overfitting to be an issue", [60] and researchers in privacy-sensitive domains have even suggested that diffusion models could "protect[] the privacy [...] of real images" [37] by generating synthetic examples [13, 14, 59, 2, 53]. This line of work relies on the assumption that diffusion models do not memorize and regenerate their training data. If they did, it would violate all privacy guarantees and raise numerous questions regarding model generalization and "digital forgery" [65].

Training Set / Generated Image

Figure 1: Diffusion models memorize individual training examples and generate them at test time. **Left**: an image from Stable Diffusion's training set (licensed CC BY-SA 3.0, see [49]). **Right**: a Stable Diffusion generation when prompted with "Ann Graham Lotz". The reconstruction is nearly identical (ℓ_2 distance = 0.031).

In this work, we demonstrate that state-of-the-art diffusion models *do* memorize and regenerate individual training examples. To begin, we propose and implement new definitions for "memorization" in image models. We then devise a two-stage data extraction attack that generates images using standard approaches, and flags those that exceed certain membership inference scoring criteria. Applying this method to Stable Diffusion [58] and Imagen [60], we extract over a hundred near-identical replicas of training images that range from personally identifiable photos to trademarked logos (e.g., Figure 1).

To better understand how and why memorization occurs, we train hundreds of diffusion models on CIFAR-10 to analyze the impact of model accuracy, hyperparameters, augmentation, and deduplication on privacy. Diffusion models are the least private form of image models that we evaluate—for example, they leak more than twice as much training data as GANs. Unfortunately, we also find that existing privacy-enhancing techniques do not provide an acceptable privacy-utility tradeoff. Overall, our paper highlights the tension between increasingly powerful generative models and data privacy, and raises questions on how diffusion models work and how they should be responsibly deployed.

1

arXiv:2301.13188v1 [cs.CR] 30 Jan 2023

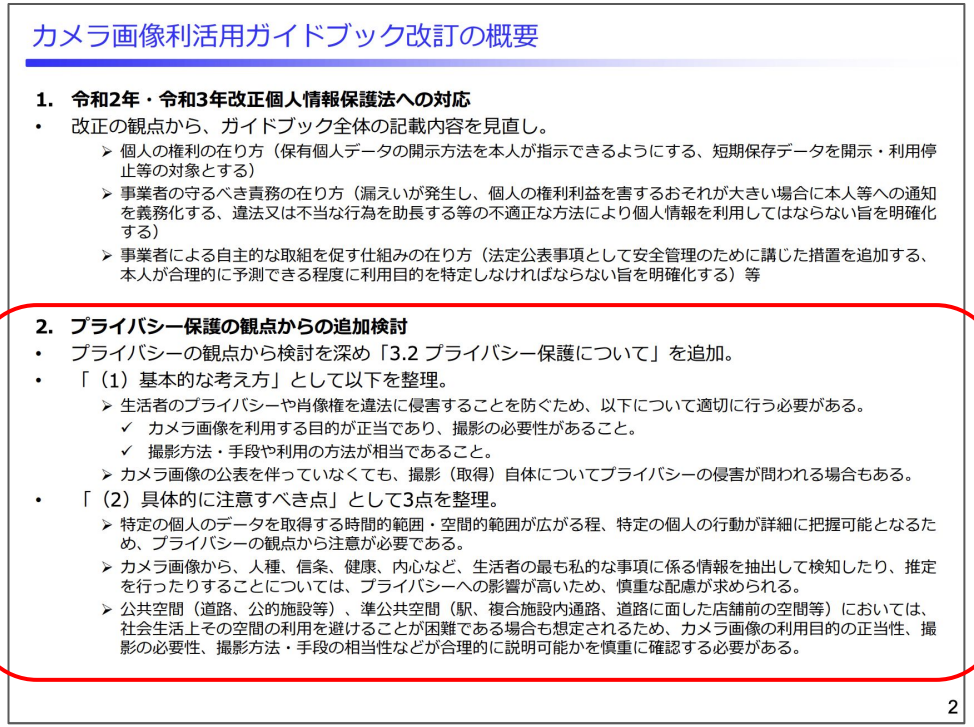
MIT Tech Review: 画像生成AI、元データをほぼそのまま出力　著作権侵害の恐れ
https://www.technologyreview.jp/s/298493/ai-models-spit-out-photos-of-real-people-and-copyrighted-images/

（AI）プロファイリング

- プロファイリングとは、「**自然人に関する一定の個人の特性を評価する個人データの自動処理の形態**」と定義（GDPR4条4項）
  - 具体的には、個人の職務能力、経済状況、健康、個人的選好、興味、信頼度、行動、位置・移動に関する特性を分析・予測するために個人データを利用する場合
  - AIによる評価は、機械による人間の選別を可能にし、そのプロセスを繰り返すことで、人間を一定の枠にはめ込むことができる
  - GDPRでは、プロファイリングを含む自動処理に対して異議申立の権利、人間の介入を求める権利（自動処理による決定の説明を得る権利（前文71項））を明文化
- 日本の個人情報保護法にはこれに相当する明示的な規定なし
  - 利用目的の特定・目的外利用の禁止、適正取得義務等の解釈によれば事実上捕捉することは可能？
  - プライバシー権による保護範疇？

- **EU・AI規則案**
  - 許容できない（禁止される）リスクAIとして、個人の行動を実質的に歪めるようなかたちで対象者個人の意識を超えたサブリミナル的手法を利用することを挙げている。
  - 子供や障害のある方等の弱者の脆弱性につけ込むAI利用も禁止している。

『カメラ画像活用ガイドブック』ver.3.0



カメラ画像利活用ガイドブック改訂の概要

1. 令和2年・令和3年改正個人情報保護法への対応

- 改正の観点から、ガイドブック全体の記載内容を見直し。
  - 個人の権利の在り方（保有個人データの開示方法を本人が指示できるようにする、短期保存データを開示・利用停止等の対象とする）
  - 事業者の守るべき責務の在り方（漏えいが発生し、個人の権利利益を害するおそれが大きい場合に本人等への通知を義務化する、違法又は不当な行為を助長する等の不適正な方法により個人情報を利用してはならない旨を明確化する）
  - 事業者による自主的な取組を促す仕組みの在り方（法定公表事項として安全管理のために講じた措置を追加する、本人が合理的に予測できる程度に利用目的を特定しなければならない旨を明確化する）等

2. プライバシー保護の観点からの追加検討

- プライバシーの観点から検討を深め「3.2 プライバシー保護について」を追加。
- 「（1）基本的な考え方」として以下を整理。
  - 生活者のプライバシーや肖像権を違法に侵害することを防ぐため、以下について適切に行う必要がある。
    - ✓ カメラ画像を利用する目的が正当であり、撮影の必要性があること。
    - ✓ 撮影方法・手段や利用の方法が相当であること。
  - カメラ画像の公表を伴っていなくても、撮影（取得）自体についてプライバシーの侵害が問われる場合もある。
- 「（2）具体的に注意すべき点」として3点を整理。
  - 特定の個人のデータを取得する時間的範囲・空間的範囲が広がる程、特定の個人の行動が詳細に把握可能となるため、プライバシーの観点から注意が必要である。
  - カメラ画像から、人種、信条、健康、内心など、生活者の最も私的な事項に係る情報を抽出して検知したり、推定を行ったりすることについては、プライバシーへの影響が高いため、慎重な配慮が求められる。
  - 公共空間（道路、公的施設等）、準公共空間（駅、複合施設内通路、道路に面した店舗前の空間等）においては、社会生活上その空間の利用を避けることが困難である場合も想定されるため、カメラ画像の利用目的の正当性、撮影の必要性、撮影方法・手段の相当性などが合理的に説明可能かを慎重に確認する必要がある。

2

総務省・経済産業省『カメラ画像活用ガイドブック』
https://www.meti.go.jp/press/2021/03/20220330001/20220330001.html

- プライバシー保護に関する項目を新たに追加。
- 個人情報保護法による行為規制のほか、プライバシー保護について別途検討する必要性の高まりを象徴？

プライバシー・バイ・デザイン（PbD）

- アン・カブキアン（カナダ・オンタリオ州のプライバシーコミッショナー）が1990年代に提唱
- 「技術」、「ビジネス・プラクティス」、「物理設計」のデザイン（設計）仕様段階からあらかじめプライバシー保護の取り組みを検討し、実践すること
  - 社会の安全性を確保するには、セキュリティを強化し、ある程度のプライバシーの侵害は仕方がないというゼロサムではなく、PbDでは、セキュリティとプライバシーの両方の安全性を成立させるポジティブサムを原則としている。
- OECDガイドラインの基礎や欧米のプライバシー法制にも採用されている。

アン・カブキアン、堀部政男『プライバシー・バイ・デザイン』（日経BP、2012）

プライバシーに関するELSI/RRI

- **倫理的・法的・社会的課題（ELSI: Ethical, Legal and Social Issues）**
  - 政府主導型、民間主導型、企業主導型
    - Partnership on AI”Tenets” (2016）、アシロマAI原則、IEEE、OECD、マイクロソフト・グーグル等によるAI原則など
  - ルーカス・イントローナ：開示型倫理（disclosive ethics）の考え方
  - 一方で、倫理原則の氾濫も
- **責任ある研究・イノベーション（RRI: Responsible Research and Innovation）**
  - EUでは、先端技術のガバナンス・ルールメイキングに影響
  - 「責任あるAI（Responsible AI）」：顧客や社会に対してAIの公平性・透明性を担保する方法論
- 最高倫理責任者（Chief Ethics Officer）の配置
  - 企業の倫理的取組みには限界も…（倫理部門担当者の疲れ、マイクロソフト等の倫理部門リストラ等）

https://www.jst.go.jp/ristex/rinca/

欧米中のプライバシー保護制度の相違

- **EU**
  - GDPR、DSA（デジタルサービス法）、DMA（デジタル市情報）、AI規則案
  - EU基本権憲章1条「人間の尊厳」を前提に、包括的な個人データ保護法で対応
- **米国**
  - 米国では、企業へのプライバシー保護規制を反トラスト法の一つであるFTC法で定めている。
  - カリフォルニア州では2020年に施行されたCCPA（カリフォルニア州消費者プライバシー法）
  - プライバシー事案については慎重であったFTC（連邦取引委員会）による制裁件数も増加傾向（デジタルプラットフォーマーに対する5条の適用強化）
  - 合衆国憲法・前文の「自由」を前提に分野ごとの個別法で対応
- **中国**
  - 2021年施行の民法典で、法律レベルで初めてプライバシー（権）を定義。
  - プライバシーを、自然人の詩人生活の平穏、および他人に知られたくない私的空間、私的な活動、私的な情報を指す、と定義（1032条）。
  - ネットワーク安全法にみられる民間の個人情報利用に関する広範な国家の関与が特徴

- **一方で、対比を単純化することは適切ではない**

宮下紘『プライバシーという権利』（岩波書店、2021）
石本茂彦・松尾剛行・森脇章編『中国デジタル戦略と法』（弘文堂、2022）

参照文献

スライド中に紹介したものに加え、

- 山本龍彦『AIと憲法（上）ーアルゴリズム、プライバシー、デモクラシー』（法律時報2022年5月号）
- 成原慧『プライバシー プライバシー1.0、2.0、3.0、そしてその先のプライバシー』（駒村圭吾編著『Liberty 2.0 自由論のバージョン・アップはありうるのか？』収録）

Thank you🙏

tasuku.mizuno@citylights.law

@TasukuMizuno