Designing Fiduciary Artificial Intelligence

Sebastian Benthall
New York University School of Law
New York, NY, USA
spb413@nyu.edu

David Shekman
Northwestern Pritzker School of Law
Chicago, IL, USA
david.shekman@law.northwestern.edu

ABSTRACT

A fiduciary is a trusted agent that has the legal duty to act with loyalty and care towards a principal that employs them. When fiduciary organizations interact with users through a digital interface, or otherwise automate their operations with artificial intelligence, they will need to design these AI systems to be compliant with their duties. This article synthesizes recent work in computer science and law to develop a procedure for designing and auditing Fiduciary AI. The designer of a Fiduciary AI should understand the context of the system, identify its principals, and assess the best interests of those principals. Then the designer must be loyal with respect to those interests, and careful in an contextually appropriate way. We connect the steps in this procedure to dimensions of Trustworthy AI, such as privacy and alignment. Fiduciary AI is a promising means to address the incompleteness of data subject's consent when interacting with complex technical systems.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Social and professional topics → Computing / technology policy**; • **Computing methodologies → Artificial intelligence**.

ACM Reference Format:
Sebastian Benthall and David Shekman. 2023. Designing Fiduciary Artificial Intelligence. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3617694.3623230

1 OVERVIEW

Fiduciary duties are some of humanity's oldest rules [42], dating back to Hammurabi's Code and legally recognized throughout the world today. Named for *fiducia*, the Latin word for trust, they concern the duties of an *agent* hired to perform a task for a *principal*. Because the agent is often in a position of expertise or power over the principal, the fiduciary duty legally bolsters the loyalty and care expected of the agent beyond the specific terms of their contract. They apply to professional roles such as health providers, legal and investment advisors, and the directors of trusts and corporations.

Fiduciary duties already apply to technology companies operating in some regulated sectors. In the European Union, under the Data Governance Act [1], data intermediaries will be fiduciaries. Legal scholars have begun exploring the possibility of expanding fiduciary duties to information controllers more generally [13, 14, 67, 105]. These broader applications of fiduciary principles have appeared in recent legislative proposals.[2]

Today, many companies engage their clients primarily through an automated user interface. When companies deliver services at scale via a website, an app, or a physical object in the Internet of Things, they delegate their operations to a computational system. These systems are increasingly trained on data from their users using general purpose machine learning algorithms and pretrained models for tasks like facial recognition and natural language processing, and hence are artificial intelligence systems[3]. Referring to automated clients as Artificial Intelligences (AI) can conflate the automated decisions made by at the interface with the ongoing choices made by engineers, content moderators, designers, and product managers. However, as the client interfaces become more automated and trained using machine learning, the more referring to them as AI becomes more literally correct.

This article is a guide to designing AI systems that are operated by people or corporations that have fiduciary duties to others that are principals or beneficiaries. We refer to an artificial intelligence that upholds fiduciary duties owed by its operator as Fiduciary AI. [4] Fiduciary AI is one form Trustworthy AI (TAI), a way of designing intelligent systems that are lawful, ethical, and technically and socially robust.

The first thing a designer of a Fiduciary AI needs is a working understanding of what fiduciary duties are and why they have been applied. Section 2 presents the legal arguments for fiduciary duties in general. Fiduciary duties are invoked when an agent holds power over its principals and when the principal is unable to completely specify the actions of the agent through consensual contract. These conditions hold in many applications of AI.

To the extent possible in an area of unsettled law, Section 3 discusses what measures should be taken by the designer of an artificial intelligence system in order for it to be compliant with these laws. We connect aspects of Fiduciary AI to known dimensions of Trustworthy AI [133, 140], including privacy and AI Alignment. We outline a procedure involving a systematic series of sensitizing questions to be asked when the AI is designed or audited for

EAAMO '23, October 30–November 01, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/10…$15.00
https://doi.org/10.1145/3617694.3623230

[1] Regulation 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), 2022 O.J. (L 152), 1-44.

[2] Federally — American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022); Data Care Act of 2021, S.919, 117th Cong. (2021). For state — H.B. 1749, 192nd Sess., (Mass. 2021); New York Consumer Privacy Act, S. 6701B, 2021-22 Sess. (NY 2021).

[3] For a standard definition of artificial intelligence systems, see OECD Recommendation on AI:2019; ISO/IEC 22989:2022

[4] This is also what we mean when we discuss fiduciary duties for computational systems, or the application of these duties to AI.

compliance. For each step in the procedure, we then identify how this aspect of Fiduciary AI has been illuminated in prior literature.

We maintain that a Fiduciary AI should be scoped to a particular context of activity (Section 4) and designed with an identified set of principals in mind (Section 5). Notably, these two steps of the procedure concern legal requirements and the prioritization of beneficiaries, which are not technical questions but nevertheless logically precede questions concerning the system's technical design.

With this contextual information in hand, the system can then be designed in technical detail. The best interest of these principals must be assessed (Section 6), which may involve machine learnt representations of the principals' best interests. If there are many principals, these interests must be aggregated (Section 7). Fiduciaries are ultimately bound to duties of loyalty and care. Fiduciary AI loyalty can be understood as alignment with the assessed best interests of the principals (Section 8). Care refers to the adherence to other best practices of community norms guiding the design of the system, such as the determination of the systems' inductive bias (Section 9).

2 FIDUCIARY PRINCIPLES AND REGULATORY TRENDS

Fiduciary duties are, foremost, legal constructions that encapsulate and enforce expectations of trustworthiness. This section discusses fiduciary law and is organized as follows. First, we will discuss fiduciary duties in general as they are understood in many sectors that predate current uses of computation and AI, with specific examples of duties that can inform Fiduciary AI design. Next we will discuss proposed expansions of fiduciary duties to new computational domains. The application of fiduciary duties to computational systems is currently an area of unsettled law. We do not in this paper take a position on which systems ought to be legally held to fiduciary principles. However, because legal outcomes are inherently normative and discursive, we present the broad legal debate to best anticipate the engineering and design principles that would guide a TAI system toward compliance.

We must note that fiduciary laws vary between jurisdictions. This paper will utilize the language of Anglophone common law to describe fiduciary relationships, principles, and actors, and focus on U.S. and European jurisdictions. The term "fiduciary duties", as it is understood in the Anglophone legal context, is difficult to translate directly across all languages. However, the duties of loyalty and care which comprise the concept have direct corollaries in the Western European context [46]. Moreover, while there are significant differences between common law and civil law variations of fiduciary duties, there has been a convergence of principles with respect to *new* fiduciary duties aimed at actors in the data economy and, by extension, AI. We find that the key principle of Fiduciary AI is that it must be aligned with the best interests of the beneficiaries in order to remedy an imbalance of power and necessary incompleteness in contracting.

2.1 Justifications of fiduciary principles

Fiduciary duties arise from a recognition that specialization is an important factor in society. It would be nonsensical and rather inefficient for every person to gain the expertise necessary to become a doctor, lawyer, or financial advisor every time they need one [42]. Instead, society relies on those individuals with whom that expertise is concentrated. However, this creates a risk for beneficiaries. Fiduciary duty dictates that fiduciaries must act with loyalty and care to their beneficiaries and face stiff penalties if they violate these duties. Fiduciary duties are justified in legal scholarship according to at least two lines of argument — an argument based on the vulnerability of the principal in fiduciary relationships, and an argument about economic efficiency under conditions when transaction costs make contracting necessarily incomplete.

2.1.1 Fiduciary relationships. The most common account of fiduciary duties focuses on the relationship between the principal and the agent, where the asymmetry of knowledge and power creates a fiduciary relationship [86]. That asymmetry comes from what Frankel [42] calls "entrustment". Patients trust doctors with their bodies; clients trust financial advisors and lawyers with their money and discretion. This entrustment of power over oneself or on one's behalf raises the potential for harm by the trustee. Because of the vulnerability this creates, additional restrictions beyond commercial transactions' standard obligation of good faith have been established to protect the principals.

Certain roles or statuses have been deemed to be fiduciaries by conventional wisdom (and common law), and fiduciary principles attach to agents by virtue of membership in such a status. Trustees, directors, agents, lawyers, and doctors are well-known examples [86]. Courts may also apply fiduciary duties in other cases wherein the agent's role bears significant enough resemblance to a fiduciary, for example because they cultivate a relationship of trust, hold themselves out as an expert in the field, are relied on by a person for advice, and have that person's complete trust [66].

2.1.2 Contractarian justification. Parallel with the fiduciary relation arguments, the Law and Economics (L&E) branch of legal scholarship has endorsed fiduciary duties on different grounds. Easterbrook and Fischel [35] leave morality out of the equation, aligning more closely with a contractarian approach. Contractarians view the entire commercial milieu, both agency relationships and the corporation itself, as composed of contracts [62]. By default, the design burden lies with the contracting parties. Easterbrook and Fischel [35] view the courts as playing an active role in specification of the contract via interim or ex post adjudication. The court functions as the theoretical completing piece of the incomplete contract between the parties [21].

This is especially important in cases where there is a severe imbalance of transaction costs against the principal party. If specifying a contract, or monitoring an agent's behavior with respect to it, is too costly for the principal, the contract will not be negotiated to completion. But the duty of loyalty (intention) and duty of care (execution) establish, in an incomplete contract, the standards that would be arrived at were full negotiation possible. The expertise of the agent is relied on to "complete" the scope of the contract according to those duties; the agent is expected to charge a premium for foregone opportunities that result from the duties. Thus the principal can engage the transaction trusting to get what they wanted from it: the agent's expertise.

Fiduciary duties are a way to solve the principal-agent problems of moral hazard and hidden action. Rather than disempower the agent, fiduciary duties allow the principal an *ex post* review of the agent's behavior. The duties are written as broadly as possible in order to cover the territory left open by incomplete contracts [115]. This enables contracts that would have otherwise failed due to the large transaction costs [35].

2.1.3 Contextuality and subsidiary duties. In the United States, the primary fiduciary duties are the *duty of loyalty* and *duty of care*. These are open-ended standards used across many sectors. In practice, these broad duties are interpreted by courts into more granular *subsidiary duties* that are "field-specific elaborations" [114] of the primary duties. The definitions of subsidiary duties crystallize the interpretation of broader duties of loyalty and care into forms that are easier to comply with and enforce. Table 1 displays subsidiary duties for a variety of professional statuses to which fiduciary duties apply, as well as speculative subsidiary duties that might apply to computational systems. In new fiduciary duties introduced by statute, subsidiary duties may be made explicit in legal clauses that accompany expressions of the more general duties. Often these subsidiary duties refer to the flow of information from the agent to or about the principal.

2.1.4 Example: the prudent investor rule. An example of a subsidiary duty is the *prudent investor rule*, a subsidiary duty of care in the context of trust law. [5] This rule, which has since been adopted by many U.S. states, aligns the fiduciary duties of a trust with modern portfolio theory [81], which is a scientifically validated standard of asset management. As a scientific expert standard, it generalizes the idiosyncratic interests of trustees in a way that is consistent with the purpose of the fiduciary as a trustee. The trustee is bound by the duty of care to this formulation of the principals' interests, while the duty of loyalty requires them to put the principal's interests over their own.

2.1.5 Example: the duty of impartiality. One example of a subsidiary duty that arises from the duty of loyalty, which will be relevant to the design of Fiduciary AI, is the *duty of impartiality*, which applies to trusts and corporate directors. This duty comes into play when there may be conflicting interests between multiple beneficiaries [50], such as trustees with conflicting preferences [110] or different classes of shareholders [91]. Impartiality does not mean that fiduciaries must treat each beneficiary equally, merely that their balancing of beneficiaries' interests is not influenced by the fiduciary's self-interest or favoritism. The corporate directors can change the weights they assign to the preference sets of differing shareholders according to circumstance [91]. We discuss this further in Section 7.

2.2 New computational fiduciaries

Recent scholarship, advocacy, and legislation has called for extending fiduciary duties to new contexts associated with AI and digital services. We briefly consider some of the arguments and controversies of these proposals here.

2.2.1 As a remedy for incomplete contracting. One key motivation for extending fiduciary duties to AI and digital services is the insufficiency of the "notice and consent" framework currently governing consumer privacy and data protection more broadly. Users of a commercial digital service sign a contract with the provider that determines the terms of their interaction, including the use of their data. It is well known that most users will not read the contracts that they digitally sign [85, 96]. In some cases, the system's design can be such that the offered consent can be considered either uninformed (c.f. [43, 106]) or manipulated [20, 23, 122]. Such a design need not be intentional. An artificially intelligent system trained to optimize an objective function (such as total clicks on advertisements) can *learn* to present an interface to users that elicits their uninformed consent. In short, the burden of the transaction costs to users of asserting their interests through contracting prohibitively high, and AI operating with conflicted interests can exploit this. Corroborating this view, Hadfield-Menell and Hadfield [56] have drawn the connection between AI Alignment and the legal challenges around incomplete contracting, and conclude that AI Alignment will depend on a larger legal framework in which a community's normative structure is imputed into the terms of the relationship between the principal and the agent. As discussed in Section 2.1.2, fiduciary duties are an available legal tool for establishing the alignment of an agent to a principal when contracting is incomplete.

2.2.2 Example: AI assistants. Aguirre et al. [5] consider specifically the case of AI assistants like Alexa, Siri, Google, and Cortana which serve as interfaces between consumers and the web. While these assistants may appear to act in the interests of their users, they may be designed with embedded conflicts of interest. Duties of loyalty can correct this.

2.2.3 Example: E.U. Data Governance Act (DGA). The European Union recognizes a fundamental right to data protection [111]. In 2016, the EU passed a landmark data protection law, the General Data Protection Regulation (GDPR), which received global attention due to its strong extraterritorially enforced sanctions [121]. Under the GDPR, consent is a legal basis for lawful processing of personal data that is often invoked in commercial applications.[6] This includes data collected about internet browsing behavior through cookies, which is a key data source underpinning targeted advertising. Consent Management Platforms (CMPs) are software solutions that collect, store, and monitor users' consent to uses of this personal data [109]. Use of these CMPs has expanded rapidly since the passing of GDPR [60]. However, these CMPs have not successfully guaranteed data protection rights as the GDPR has intended. [7]

[5] Originally, the "prudent man rule" was a common law standard enshrined in 1830 (*Harvard College v. Amory*) that a trustee prioritize regular income over speculative value when managing trust assets. Later, this was updated in the American Law Institute's Restatement Third of Trusts and promulgated as the Uniform Prudent Investor Act (UPIA). (Restatement (Third) of Trusts § 90 (2007), 61 A.L.R.7th Art. 1)

[6] Article 6 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 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), art. 6, OJ 2016 L 119/1.

[7] The Transparency and Consent Framework (TCF) [38], a standardized CMP developed by IAB Europe (an association for the digital marketing industry), was found to violate GDPR by Belgium's Data Protection Authority on several grounds. (Belgian Data Protection Authority, Decision on the Merits 21/2022 of 2 February 2022, Complaint Relating to Transparency & Consent Framework (IAB Europe), DOS-2019-01377, https://www.autoriteprotectiondonnees.be/publications/decision-quant-

Context	Loyalty	Care	Both
Trusts	Incurring only reasonable costs	Prudent investor rule Duty to not commingle trust property	Giving account to beneficiaries† Record-keeping†
Corporate management	No usurpation of corporate opportunity No impairment of shareholder meetings Boardroom confidentiality†	Need for monitoring and compliance	Disclosure to shareholders†
Investment advice	Best execution of instructions	Prudent investor rule	Keeping books and records†
Legal representation	No conflicts of interest	Safeguarding the client's confidences†	Communication with the client†
Health care	Confidentiality†		Informed consent†
(DGA) Data Intermediaries‡	Facilitate exercise of rights Act in subject's best interest	Security and confidentiality† Ensure interoperability	Notice of data uses† Consent management
Data [59] Processors‡			
Collection	Data minimization	Proper disposal†	Data records†
Personalization	No conflict with collective best interests	Eliminate disparate impact	
Gatekeeping	No exposing principals to privacy harms	Counterparty evaluation	Proper security practices†
Influence	Eliminate dark patterns		
Mediation	Disincentivize user harms Hierarchical loyalty for heterogeneous roles	Mediation algorithms oversight	
AI [5] Assistants‡	No conflicts of interest Transparent objective functions	Clearly indicate potential conflicts† Adequate requests of user input†	

Table 1: Examples of subsidiary duties by legal field. Includes hypothetical examples of subsidiary duties for information processors. Subsidiary duties marked with a dagger (†) concern information flows. Contexts marked with a double dagger (‡) are fiduciary contexts not yet determined by law but suggested in the literature [5, 59].

European lawmakers have passed the Data Governance Act (DGA), which addresses the uncertain legal status of CMPs by identifying them as a type of "data intermediary". [8] According to the DGA, "data intermediation services providers seek to enhance the agency of data subjects, and in particular individuals' control over data relating to them".[9] The Act uses the language of fiduciary duties and "act[ing] in the best interests of the data subjects".[10]

au-fond-n-21-2022-english.pdf) Others have frustrated privacy scholars and users alike by employing dark patterns and choice architecture to push users toward invalid or uninformed consent [95]. Indeed, researchers have found that in widely deployed CMPs, data is collected, processed and shared even when users have not consented to it [80].

[8] *See* Recital 30, Regulation 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), 2022 O.J. (L 152), 11.

[9] Recital 30, Regulation 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), 2022 O.J. (L 152), 11.

[10] "Data intermediation services providers that intermediate the exchange of data between individuals as data subjects and legal persons as data users should, in addition, bear fiduciary duty towards the individuals, to ensure that they act in the best interest of the data subjects." Recital 33, Regulation 2022/868 of the European Parliament and

Under this law, data intermediation is its own fiduciary context. The DGA restricts data intermediaries to using collected data for only the purposes of its intermediation services. It further prohibits the pricing and other commercial terms of the service from being dependent on the customer's degree of use of any other service. In short, data intermediaries under the DGA are fiduciaries that assist individuals in exercising their data rights for their best interests.

2.2.4 Information fiduciaries and their discontents. While fiduciary principles are already relevant to computational systems operated in sectors regulated by existing fiduciary law, and to data intermediaries by the DGA, recent work has proposed the expansion of fiduciary principles to data processors more general, creating so-called "information fiduciaries" [13] with "data loyalty" [59]. Several bills have been proposed both at the federal and state level aiming to apply duties of loyalty and care, as well as subsidiary duties, to platforms.[11] [12]

Broad information fiduciary laws are controversial. They would upset how many well-established multi-sided platform companies [29], which necessarily navigate conflicting interests of multiple stakeholders, do business – and hence how they would use AI, such as recommendation systems. This raises legal questions about how an information fiduciary would reconcile duties to shareholders with conflicting duties to principals in other roles, such as users and advertising partners [54, 67]. However, corporate fiduciaries are still bound by the measures of the law [127]. Committing an illegal act would be detrimental to the financial value of the company (either through impositions of fines or other penalties, reduced revenues, or loss of brand value/goodwill) and thus the shareholders' interests. If fiduciary duties to users of platforms were enacted into law, then that would be another legal constraint within which the platforms' corporate officers would need to act. We generalize from this example and consider the prioritization of multiple beneficiary roles in Sections 5 and 8.

3 ENGINEERING FIDUCIARY AI SYSTEMS

For the reasons discussed in Section 2, many technology businesses may find that they have fiduciary duties. Given that businesses increasingly rely on automation and AI for their operations, and that they can be accountable to fiduciary duties for these operations, these businesses should be aware of how their automated systems can be designed *ex ante* to avoid violation of their legal obligations. To this end, some technical interpretation is helpful for identifying what constitutes compliance with the law and to what extent compliance can be guaranteed by an auditor. [13]

The remainder of this paper assumes that a technologist is interested in implementing a fiduciary service or artificial agent. What do fiduciary duties mean for their technical practice and design? We contribute a procedure for a system architect to consult when attempting to build a new system that is compliant-by-design with fiduciary duties, or when auditing an existing system for compliance. It is formulated as a rubric of questions. (See Table 2.) The questions are presented in a logical sequence such that the answer to early questions, such as "What is the context of the system?" and "Who are the principals?" inform the answers to later questions, such as "Is the system aligned with the best interests of the principal?"

While many of the designers decisions are quite technical, we find that subsidiary fiduciary duties may directly inform the design at many stages. For example, a reasonable person standard from a subsidiary fiduciary duty can inform the answer to the question, "What are the best interests of the principals?", which would otherwise be derived from data. In this way, subsidiary duties may guide *ex ante* expectations of behavior and guide designs. However, fiduciaries will also be held to the more abstract, primary fiduciary duties of loyalty and care. These latter duties allow for flexibility by the courts and regulators in *ex post* analysis.

Further sections of this paper address facets of our rubric for Fiduciary AI by outlining corresponding techniques or frameworks from computer science. Naturally, in an area of unsettled law, using such a rubric will be necessary at best, and never sufficient, for compliance with regulations.[14] And in many cases, the TAI dimensions implicated by fiduciary duties are active research areas with many open questions. We see Fiduciary AI as an emerging area of both law and engineering. What follows is a survey of the state of the art at the intersection of these fields.

We see Fiduciary AI practice as involving a novel combination of Contextual Integrity [92], a theory of ethical computing design that focuses on contextualized social norms, and the problem of AI Alignment. The AI alignment problem, or how to design an AI that acts in a way that is aligned with the intentions of principal users, has been most widely discussed in the growing field of Artificial General Intelligence (AGI) research [31, 40, 58]. Fiduciary duties are one form of legally binding, contextualized normative structure that can address the necessary incompleteness in the relationship between AI and its users [56]. Thus, Fiduciary AI design can be a precursor to a better understanding of Aligned AI [5].

4 CONTEXT

In the law, there are no absolute fiduciaries who are required under all circumstances to serve the interests of a principal. Rather, individuals are fiduciaries to others by virtue of their respective

of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), art. 12(m), 2022 O.J. (L 152), 12-13.

[11] Federally — American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022); Data Care Act of 2021, S.919, 117th Cong. (2021). For state — H.B. 1749, 192nd Sess., (Mass. 2021); New York Consumer Privacy Act, S. 6701B, 2021-22 Sess. (NY 2021).

[12] This trend is not limited solely to the US. In India, the Digital Personal Data Protection Act, 2023, explicitly creates data fiduciaries and data principals. Though the language of data fiduciaries is used in this bill, the functional implications appear to align closer with that of the "data controller" under the GDPR regime.

[13] Of course, it may be that no *ex ante* design will be sufficient for compliance with fiduciary duties, given the possible *ex post* actions of businesses and the variability of judicial interpretation. Some legal scholars are openly skeptical about the possibility of engineering compliance [27, 137, 138]. However, we will connect fiduciary duties to Trustworthy AI principles and techniques as a pragmatic exercise. We acknowledge that there are both limits to [112] and value in [2] the use of computational abstraction to clarify normative guidelines. But for AI systems in particular, designers *must* make ethical choices. For example, Baum [15] argues that there are three sets of decisions in a social AI system that require the imposition of the designer's values: standing, measurement, and aggregation. We find that analogous questions arise when designing a fiduciary AI. We have replaced "standing" with "identification", and "measurement" with "assessment". Standing, in Baum's context, is "to have one's ethics included in a social choice process used to determine the ethics of AI" Baum [15]. However, this differs from the use of the term "standing" in law to refer to whether one has the ability to bring suit.

[14] In the future, such sufficient conditions could be defined by law via a "safe harbor" provision, which is essentially a well-understood floor of expectations above which a covered entity will be considered to be compliant.

1. Context	What social context or sector will the system will operate in?	Section 4
	What are the purposes of that context?	
	What roles are defined in that context?	
	What norms or rules does that context imply?	
2. Identification	Who are the principals?	Section 5
	Is there more than one category of principals?	
	Do the categories have different levels of priority?	
3. Assessment	How are the interests of the principals being assessed?	Section 6
	Are they being empirically measured? With what data?	Section 6.1
	Are future rewards discounted? By how much?	Section 6.2
	Is there are reasonable person standard of best interest?	Section 6.3
4. Aggregation	If there is more than one principal, how are their interests aggregated?	Section 7
	Whose interests dominate in a conflict?	
	Are there duties about aggregation, e.g. impartiality?	Section 7.2
5. Loyalty	Is the system aligned with the principals' interests?	Section 8
	Have subsidiary duties of loyalty been considered?	Section 8.2
6. Care	What is the context-appropriate level of prudence?	Section 9
	Is the inductive bias of the system up to contextual standards?	Section 9.1

Table 2: Steps for designing or auditing a Fiduciary AI for compliance, with corresponding sections of this paper where they are discussed.

roles in a legally recognized context. A first design principle for Fiduciary AI is that the context within which the AI is acting as a fiduciary must be established and understood.

The study of how context should inform computational system design has deep roots in ubiquitous computing research [3, 32, 34]. We recommend that Fiduciary AI draw on Nissenbaum's Contextual Integrity (CI) [92], a well-developed theory of social contexts that is used in the design and evaluation of technical systems[15]. It begins from the position that a pluralistic society [139] can be understood to consist of multiple social contexts or spheres, each with its own internal system of roles, rules, and meanings. CI offers a rubric for understanding contexts in terms of:

- *Purposes.* The purpose of the social context in society, towards which its laws and norms are instrumental.
- *Roles.* The names and definitions of roles that agents can have in that context.
- *Norms.* Social or legal rules prescribed within the context about how agents of the identified roles should interact.

Translating the CI framework into technical requirements is an ongoing area of computer science research [17] We do not recommend any particular implementation of CI, but rather recommend using it to frame the answer to a specific question: in what context is an AI expected to be performing fiduciary duties? This understanding of context in terms of roles, purposes, and norms will be used to answer other questions related to identification (Section 5) and aggregation (Section 7). Furthermore, CI is a fitting framework for evaluating the context of fiduciary duties because some subsidiary duties, like duties of disclosure and confidentiality, are explicitly about information flow. We develop this case as an example of alignment in Section 8.2.

5 IDENTIFICATION

Designing Fiduciary AI requires an explicit understanding of who a system's principals are. The question "who are the beneficiaries of this system?" has long been one of the first questions to ask when designing any social system [130]. In most cases today, AI is designed mainly to benefit those that deploy the system, using mechanism design techniques to steer the behavior of users towards those goals [136]. Indeed, some of the literature on AI alignment assumes that the goal of AI safety is to ensure that the AI system is aligned with the interests of its operator [6], while others defer the question of the "preference payload" to other research [77].

Fiduciary AI can be designed with other categories of principals in mind, and identifying these principals is essential for compliance with fiduciary duties. This choice cannot be founded on purely technical considerations; rather, it must be inferred from the previously identified legal or social context (see Section 4). For example, many companies that build AI products operate under Delaware corporate law and their corporate directors have only one pressing set of fiduciary obligations: to their shareholders. But a business that is based in Delaware but also is operating as a data intermediary in the European Union would need to prioritize its duties to its data

[15] Contextual Integrity is mainly about norms of information flow, and defines privacy as appropriate information flow. Each norm is parameterized in terms of actors (sender, receiver, and data subject); attributes (meanings of information) and transmission principles (for instance confidentiality, reciprocity, or "with a warrant"). The analyst is encouraged to evaluate the introduction of a new technology with a heuristic procedure that identifies how any new information flows depart from the norms embedded in the context of interest. One of the steps in the procedure is to evaluate any changing information flows in terms of the purposes of the context as it functions in society more broadly. Ultimately, the norms are animated by the specific purpose of that context, as well as societal values and individual goals.

subjects because legal penalties for violating those duties would be, in turn, bad for the shareholders (see Section 2.2.4).

To illustrate the significance of the fiduciary relationship, we consider a typical online system that interacts with human users in a commercial context. In the absence of a fiduciary relationship, the human user's engagement with the system will be regulated by a contractual agreement to which the user has legally consented. (See Section 2.2.1.) We note that *for those users identified as principals of a fiduciary system*, consent is *not sufficient* for the fulfillment of the fiduciary duty. Rather, other requirements, such as loyalty (see Section 8), must govern the interface design to ensure as much as possible that consent, when given, is in the interests of the principal.

In a similar vein, consider a user who has consented to a system's broad terms and conditions and now operates the system through a user interface. Normally, a system would be expected to perform whatever actions the user chooses for it to do. However, for a complex system, the instructions given will never be a complete description of the operations to be performed, and the choice of instructions may be manipulated by a system's design. Indeed, there is a division in law between two models for the interpretation of the duty of loyalty — the *obedience* model, wherein the fiduciary must follow directions given by the beneficiary, and the *best-interests model*, wherein the fiduciary acts with more discretion but in service to the beneficiary. The latter best-interests model is favored by Richards and Hartzog [105] for privacy and information platform cases, as it is precisely the complexity of the technical systems and inability of consumers to grasp their mechanics which makes contracts (as well as instructions) incomplete or impossible to monitor. This requires judicious design choices on the part of the data controller, e.g., the choice of an email provider to maintain deleted emails on record for a short period of time in case the user ordered their deletion by accident.

In summary, while most digital or AI systems will be operated by users on the basis of consent and obedience, Fiduciary AI imposes a higher standard on the system's design with respect to its *principal* users. For example, a medical AI might be designed to serve the best interests of its medical patients (the principals), but have a different interface used by system operators (not principals) that is more strictly obedient. For the remainder of this article, we will assume that the principals have been identified and the Fiduciary AI designer must now determine how to fulfill the system's duties to those principals.

6 ASSESSMENT

Once the principals of a fiduciary AI system have been identified, system designers should assess the salient best interests of those principals. This is not a trivial problem by any means: how many people can judge what is in another person's best interest, or even their own? Luckily, this problem is constrained; the system designer must only assess the best interests of the principal in their role within the fiduciary context.

One approach to assessing these best interests is to learn an objective function from data provided by the population of principals. The data may be observations of behavior, statements of preferences, or direct measurements of user well-being. We will discuss these data sources and their associated challenges.

Beyond empirical assessments of best interest, the law also furnishes some fiduciary contexts with *reasonable person standards*, which are legal codifications of the interests of a normal or general person in a given role. In Fiduciary AI cases, a combination of empirical data collection, machine learning, and reasonable person standards may be used to determine principal best interests.[16] Once the information about each principal's best interests are identified, they can be aggregated if appropriate (Section 7), and then rendered into technical design (Section 8), perhaps as an objective function or reward model [77].

6.1 Learning reward models

One strategy for assessing the best interests of principals is to attempt to learn a proxy model of those interests from observational data. This approach has been best explored in the context of reinforcement learning. While we do not limit the scope of this work to reinforcement learning, we nevertheless point to reward modeling [77] or reward design [57] as key elements of Fiduciary AI.

Consider the Bellman equation form of the intertemporal decision problem:

$$V(s) = \max_a r(s,a) + \beta \mathbb{E}\left[V(s')|s,a\right] \tag{1}$$

Where V is the value function for a state s, given an optimal action a. An implied probability distribution $P(s'|s,a)$ governs the state transitions. The value function is defined recursively as the sum of immediate rewards $r(s,a)$ and the expected value of future rewards $\mathbb{E}\left[V(s')|s,a\right]$, discounted by a factor of $0 < \beta < 1$.

We define Q as the state-action value function:

$$Q(s,a) = r(s,a) + \beta \mathbb{E}\left[V(s')|s,a\right] \tag{2}$$

An agent's policy π maps from states to actions. In the reinforcement learning context, the policy is trained to optimize the action-value function.

$$\pi^*(s) = \arg\max_a Q_\pi(s,a) \tag{3}$$

Researchers working in AI Alignment have worked on techniques for learning a proxy reward function r that reflects the true objectives of a system's principals. This function can then be used in training the AI system's policy through reinforcement learning.

Reward learning comes with many challenges. In all forms of reward design, the proxy reward function will, at best, be based on observations of the true reward function, leaving the possibility for misalignment due to incompleteness [145]. It is likely that the best approach will combine several different, complementary forms of human feedback, perhaps guided by a common framework [63]. Online reward learning, wherein the reward function is continuously trained on data learned through its operation, opens the learning process up to manipulation by the system during training [7, 41], so it is best to be avoided. Furthermore, each data source has associated threats to its validity as a measure of best interests. And in all cases,

[16] While some researchers have found improved results using observed human behavior or feedback to train a policy function directly instead of training a reward function [53, 70], we do not recommend this as a way of achieving Fiduciary AI, because Fiduciary AI is specifically concerned with the best interests of the principals. We also do not recommend, for this application, inferring reward functions from observable states of the world [113], as Fiduciary AI has been proposed as a solution to problems with the status quo due to, for example, market failure.

the inference of the proxy function will have inductive bias [141] (see Section 9.1).

6.1.1 Observations of behavior. Learning the objective function of an agent based on observations of their behavior is an enduring research problem. One early formulation of this problem is the *inverse reinforcement learning* (IRL) paradigm [1, 90, 146] . In the simplest formulations of inverse reinforcement learning, a subject's behavior is observed and this is interpreted as a policy π^* that corresponds to their value function V. In principle, this allows the analyst to derive the subject's reward function r.

IRL faces a number of challenges. Ng et al. [90]'s original analysis showed that IRL suffers from the problem of *degeneracy* – that there are many reward functions for which an observed policy is optimal. Gleave et al. [49] have shown that reward functions can be grouped into equivalence classes based on their effects on policy, which helps quantify distances in the space of reward functions. But in any case, the outcome of IRL will depend on inductive bias (Section 9.1) as well as assumptions about temporal discounting (see Section 6.2).

Another set of challenges concerns how well the behavior of users reflects their best interests. Indeed, one might want a fiduciary service precisely when one is *unable* to act (or represent) in one's best interest. And as a matter of general fact, people do not always act in their best interest due to cognitive biases [64]. While there have been some efforts to adapt IRL to a "two system" cognitive architecture [101], it has been shown that, in general, reward functions in an IRL context are unidentifiable if the principal's behavior is irrational [9].

A promising direction for IRL is Cooperative Inverse Reinforcement Learning, wherein the human trainer is an expert who deliberately teaches the system its objectives [58]. In other words, the system is not trained on behavioral data from "in the wild". Behavioral data, and especially expert training data, may be be most useful in practice when combined with other sources.

6.1.2 Preference judgments. Another way to gather data about principal's objectives is to solicit their explicit preference judgments. Noothigattu et al. [94] design a system for learning "trolley problem" ethics by collecting data from human users about their preferences over of different outcomes. In the context of deep reinforcement learning, objective functions have been trained from human judgements about the value of fragments of the system's behavioral trajectories [26]. These fragmentary judgments are then aggregated and recomposed into a complete objective function. Preference judgments can be fruitfully complemented with other sources of information such as expert demonstrations [61]. Researchers have experimented with grounding natural language commands into the reward function of agents in an IRL context [12, 44, 119].

A challenge associated with using this data source is that human preferences do not have the normative properties of associated with mathematical decision-theory, such as transitivity and insensitivity to alternatives [129]. They are also unstable; well-known cognitive science results have demonstrated that human preference judgments are susceptible to framing effects, especially framing of a decision as a chance of gain versus a risk of loss [128]. Social scientists who survey the preferences of broad populations have learned that for studies to be effective, subjects need to have a positive reason for valuing the accuracy of the result [144]. In a rare treatment of these threats to validity in the AI training setting, Thomaz and Breazeal [124] find that outcomes are sensitive to how human trainers understand their relationship with the learning robots. More research is needed to understand the human side of learning reward functions from human preferences.

6.1.3 Measuring well-being. Another way to discover principal preferences is to measure their well-being directly. Behavioral economists such as Kahneman and Krueger [64] have addressed the problem that actions do not reveal true preferences and have developed alternative, independent methods of measuring well-being, such as survey techniques. Well-being measurements like these have been employed in psychology and human-computer interaction (HCI) studies of the users of social media platforms, and used to assess user interface designs [22, 72, 99]. IEEE [28] has since published scientifically valid well-being metrics and a Well-Being Impact Assessment (WIA) process.

Well-being metrics are sometimes gathered to inform public policy [51]. They can combine both objective elements (such as health or employment) [79] and subjective elements (such as frequency and duration of positive feelings) [33]. Well-being measurement is therefore largely separate from the fiduciary context and may be a noisy signal with respect to the contextualized objectives of the principal. Nevertheless, they may be a worthwhile complement to behavioral and preferential data.

6.2 Temporal discounting and time inconsistency of preferences

Temporal discounting presents further complications and opportunities in the assessment of principal interests. In inverse reinforcement learning, the discount rate β is a free parameter whose value must be either assumed or fit when deriving a human model's reward function from their behavior.[17] Rothkopf and Dimitrakakis [107] explore techniques for inferring discount factors in IRL through a Bayesian process.

However, time inconsistency is yet another way in which principals may be irrational, frustrating attempts to learn their "true" objective function. Empirically, people's discounting of future utility takes the form of hyperbolic discounting, as opposed to exponential discounting [18, 25, 52].[18] This means that even if a discount factor is fit to human behavior or preference judgements, it may an artifact of the method than a representation of the human subject's psychologically held values. Correcting these inconsistencies may be a positive role for a fiduciary to play as an advisor. Puaschunder [104] argues that a data fiduciary may be duty-bound to correct

[17] Mathematically, the discount factor $\beta < 1$ is necessary for the sum of the infinite series of expected future rewards to converge on a finite value.

[18] Hyperbolic discounting means that delays that are closer in time feel more costly than the same delays further away in time, such that, for example, one prefers to pay more to prevent a delay from today to tomorrow than they would to prevent a one-day delay a year from now (from 365 days from today to 366 days from today). Hyperbolic discounting and closely related ideas of time inconsistency of preferences and present bias [76, 97] may help explain the "privacy paradox," wherein a concern for future privacy risks can be easily subsumed in the present by a desire for instant gratification [4]. We do not use the phrase "privacy paradox" uncritically. A robust discussion of the privacy paradox is beyond the scope of this paper. C.f. [118]

these inconsistencies on behalf of their principals and be more rational, or even more patient, than they are.[19]

6.3 Legal constraints

Sometimes a legal rule can define a context-specific sense of "best interest" that simplifies assessment a great deal. For example, the prudent investor rule (see Section 2.1.4) established that for the management of trusts, the principal's interests are best served by investing according to modern portfolio theory. This simplifies the fiduciary's role to one of maximizing risk adjusted returns on investment; they do not need to take into account non-monetary measures of well-being. An information fiduciary statute could likewise construct a standard of user interests based on a combination of a strictly defined context and scientific results.

While such standards simplify the assessment process, they may allow system designers to ignore important differences in user preferences, wheres privacy preferences have been shown to be quite heterogeneous [37, 84, 131]. (An alternative approach would be to assess preferences individually, and then simplify the fiduciary's goal function through aggregation, as discussed in Section 7.2).

Translating these legal constraints into a technical specification is a challenge in its own right. In some cases, a translation into a formal specification language such as linear temporal logic is possible [30]. If the rule is better operationalized as a component of or constraint on an objective function, it may need to be trained based on expert-provided data. For example, Noothigattu et al. [93] have used IRL to learn ethical constraints for a system that is otherwise trained with reinforcement learning.

7 AGGREGATION

As discussed in Section 2.1.5, one of the conceptual challenges raised by computational fiduciaries is how a system can be loyal to the interests of multiple, potentially conflicting principals. A natural strategy is to aggregate the many objective functions assessed for each principle into a single objective function [94]. However, this strategy invites criticism from the field of social choice and voting theory, which has long raised difficulties with preference aggregation. We survey these critiques and the proposed remedies for social choice in AI, which involved partially ordered objective functions.

The available legal logic for managing conflicting fiduciaries [50, 110] provides additional constraints on the aggregation function that may simplify the problem. We consider subsidiary duties of impartiality and the prudent investor rule as examples of how law can guide aggregation.

7.1 Impossibility theorems

Social choice, the field that theorizes how to combine many individual preferences into a collective decision, has revealed many theorems that show that it is impossible for such mechanisms to meet all of several theoretically desirable criteria. A prominent example is the Gibbard-Stratherthwaite theorem[47], which states that for any ordinal voting system with one winner, the rule either:

- is dictatorial, meaning that there is one voter who can choose the winner, or
- limits the outcomes to only two possibilities, or
- is susceptible to manipulation through insincere ballots.

Manipulation becomes a problem when the assessment of best interests is based on the principals' own voluntary actions or expressions. These may be untrue or misleading [41]. Putting aside the question of incentive compatible mechanism design, we can note that manipulation may be avoided by decoupling the assessment process from aggregation. In other words, this is a reason to avoid doing best interest assessment via on-line learning [7]. More broadly, this and other impossibility theories make the selection of an aggregation rule a thorny problem. Baum [15] argues that such aggregation rules are necessarily moral choices of the designer of a social AI. Aggregating the best interests of multiple principals of a Fiduciary AI is a structurally similar task.

Other solutions to voting paradoxes involve a relaxation of the requirement that preferences or objectives be given as a total ordering. Prasad [103] proposes that for the highest hierarchical levels of decision-making in AI design, it is best to use an approval voting mechanism, which gives each voter a binary choice of consent for every option, rather than an ordinal preference. This form of voting is less prone to manipulation. It takes as its goal "consent maximization", as opposed to "utility maximization". This is not to be confused with the initial consent to use the fiduciary service, but rather is meant to be a heuristic for aggregating the interests of those already identified as principals. Eckersley [36] addresses the impossibility theorems of Arrhenius [11] and [100] in population ethics which show that there is no way to develop a totally ordered objective function over outcomes for a population without violating human ethical intuitions in one of several ways. This problem can be avoided if the aggregate objective function is a partially ordered, as opposed to totally ordered. Indeed, perhaps the most straightforward way to aggregate individual objective functions is using multi-objective optimization techniques that reveal Pareto efficient indifference curves over individual welfare. [55, 82]

In sum, the design of the aggregation function is an important choice with ethical implications. The relaxation of total ordering conditions eases some of the design burden. For the remainder of this article, we assume that objective and utility functions may be partially ordered.

7.2 Legal constraints

The problem of conflicting principal interests has arisen for non-computational fiduciaries, and in some cases it has been settled by further legal rules. These tend to be context-specific subsidiary duties (see Section 2.1.3). We consider two examples here which may be guides to how subsidiary duties could inform Fiduciary AI in the future.

In some contexts where there are potentially conflicting principals, the duty of loyalty can include a subsidiary *duty of impartiality*. As discussed in Section 2.1.5, impartiality requires that the system

[19] The financial context suggests how patience might be a virtue of fiduciary advice. Differences in discount factor are one macroeconomic explanation for the varied distribution of wealth in society [24, 74]. An impatient consumer is more likely to consume more today and save less for the future. One beneficial role of an investment advisor fiduciary is to nudge their client to consider the long-term implications of their decisions, in effect advising the principal on what they should do if they discount their future less. A Fiduciary AI might learn a principal's reward function R assuming one level of patience, and then provide recommendations to them based on the same R, but a greater patience β.

not be unduly influenced by either self-interest (of the agent) or favoritism (towards one of the principals). This constrains the agent, but the duty of impartiality also gives the agent the flexibility to act with discretion within those constraints. Just as corporate directors can make decisions that unequally impact different classes of shareholders [91], an information fiduciary could make decisions that unequally impact different classes of users without violating their duties. As a further guide to how to balance principal interests, the agent should look to the specific purpose for which the relationship with the beneficiary was established.[20] So in these cases, the agent's aggregation problem is constrained in some ways, but relaxed in others.[21]

Other legal constraints can more directly specify the agent's aggregation function. Consider again the prudent investor rule (see Section 2.1.4), which defines a principal's interests as those identified by modern portfolio theory. This ruling irons over what might otherwise be wrinkles due to idiosyncratic principal preferences, because those preferences may not be prudent. So, for example, if one principal wanted their trustee to divest from certain businesses for political reasons, it would not necessarily be the trustee's duty to do so. An information fiduciary statute could likewise construct a standard of user interests based on a combination of a strict definition of context and scientific results.

8 LOYALTY AND ALIGNMENT

Loyalty is one of the two pillars of fiduciary duty. Following Richards and Hartzog [105], we focus on a "best interests" interpretation of loyalty for Fiduciary AI because instructions given to an AI will be incomplete, leaving open questions of its behavior [56]. We now assume that through the proceeding steps of the Fiduciary AI procedure, the designer has in hand an objective function (perhaps only partially ordered) that represents the assessed and aggregated best interests of the principals. In this section, we discuss guarantees of loyalty in more depth. In particular, we note the legal idea of the "loyalty two-step": the combination of a general duty of alignment, and contextually specific subsidiary rules with clearer requirements [59]. The general duty of loyalty bears some resemblance to the broader issue of AI Alignment, and can be characterized by optimization of a proxy objective function. Subsidiary duties of loyalty can provide firmer constraints on system behavior, and address problems that are difficult to solve using machine learning.

8.1 Best interest and value alignment

The first step of the "loyalty two-step" is a "no-conflict" rule in the system design [59]. This means that the system must not be designed to conflict with the best interests of the principals. The designer must use their representation fo the best interests of the principals in good faith.

When the behavior of the system is an AI, the system can be trained to optimize this proxy objective function. In the reinforcement learning context, for example, the objective function can be used as the reward function with which the agent's policy is trained [77]. If there are hierarchical fiduciary duties between multiple classes of principals, lexicographic [120] or hierarchical [98] methods in multi-objective optimization can be used to maintain this prioritization.

In the broader research program of AI Alignment [31, 40, 58], it is understood that alignment through proxy objective functions can fail due to errors. Leike et al. [78] distinguish between the *specification* problem – when the proxy function is different from the "true" principal objectives – and the *robustness* problem – when the proxy function is accurate, but other problems in the implementation of the system lead to misaligned behavior. The requirements of specification are uncompromising; Zhuang and Hadfield-Menell [145] demonstrate that incompleteness – under-specification – of the proxy objective function can lead to arbitrarily costly behavior of the agent. Robustness problems arise not from misalignment but from other forms of error in system calibration, such as distributional shift. In Section 9, we suggest that mediating these problems be understood as a requirement of the duty of care.

Aguirre et al. [5] and others have drawn a connection between the alignment implied by a duty of loyalty and what is studied more broadly as AGI Alignment. We maintain that Fiduciary AI is sufficient, but not necessary, for AI Alignment. Fiduciary AI is a stricter demand than mere alignment because of the constraints from contextual scope. For example, when *data minimization* – the requirement that data be collected and stored only when necessary to pursue legitimate purposes – is a subsidiary duty of loyalty, this limits the power of an AI to act beyond its intended design. An AGI that is well designed as a Fiduciary AI would, as a matter of its own duties, limit the vulnerability of its principals by performing only within its narrowly defined role [59].[22]

8.2 Subsidiary duties of loyalty

When a designer schematizes the context of a Fiduciary AI system, they should enumerate any context-specific norms and rules (Section 4). These include subsidiary duties of loyalty, the second part of the "loyalty two-step", and can include other TAI standards such as: data minimization, nondiscriminatory targeting, reducing data sharing with third parties, absence of dark patterns, and content moderation [59]. (See Table 1.)

Some subsidiary duties may reference the best interests of the principals, but in more nuanced ways than broad "alignment". For

[20] This resolution to the multi-principal system design runs parallel to the way CI theorizes the formation of information norms: as best fit to a balance of the role-based ends of actors and the purpose of the social context.

[21] The Trustworthy AI dimensions of fairness and privacy are both relevant to aggregation in ways that are not currently considered in fiduciary law. In privacy scholarship, there is a growing recognition of group privacy [88, 123], when a group of people has a collective stake in their data, perhaps because of horizontal data effects [135], meaning when the data collected from one subject has implications for the interest of other subjects. There is an opportunity when designing the aggregation function to acknowledge these externalities or relational concerns and balance the weighting of preferences accordingly. Particularly well-studied in the Trustworthy AI context are fairness criteria. These fairness metrics are primarily aimed at preventing group-based discrimination in machine learning systems. However, a Fiduciary AI designer concerned with compliance with nondiscriminatory regulations or social expectations could perhaps adapt these fairness criteria to the aggregation function as well. We pose this as a problem as one for future work.

[22] Perhaps there is one exception. An AI that served as a corporate director, with fiduciary duties to its shareholders, may require general intelligence to operate many aspects of the corporation, even though their duties are narrowly construed in terms of protecting the value of the shareholder's investment in the corporation. Indeed, this sweeping scope of AI power, combined with the narrowness of its objective function, makes this form of AI particularly risky [16, 108].

example, the *duty of disclosure* is a positive rule mandating that the fiduciary reveal relevant facts to the beneficiary, especially if these are facts pertaining to a potential conflict of interest. This duty creates an opportunity for the principal to contest an action that is in potential conflict.[23] Fiduciary AI designers might engage such a duty when, for example, a system identifies that its objective function is incomplete with respect to its current decision or situation [145], and more information from the principals are needed to guarantee a lack of conflict. Inversely, the duty of confidentiality is a negative rule prohibiting the disclosure of information (typically, about the beneficiary) for the benefit of the fiduciary. This subsidiary duty involves representations of both the principal's interests *as well as* the agent's "true" interests, such as the interests of its shareholders.

Notably, these subsidiary duties govern flows of information to and from the agent. They can be understood as transmission principles — norms governing flows of information as per Contextual Integrity [92]. Formal work on the strategic value of information and its import for AI alignment has been pursued by Everitt et al. [39].

9 CARE

Some jurisdictions and contexts define a fiduciary duty of care. The duty of care obliges the agent to prevent any foreseeable harms to the principal. It specifically holds the agent liable for accidental harms to a higher degree than they would be held otherwise under tort law. Whereas loyalty requires the agent to make decisions in the best interest of the principal, care requires the agent to be *highly informed* before making decisions, so as not to make a decision recklessly.

Fiduciary agents are held to a standard of prudence that is defined contextually. For example, the executor of a trust is held to the prudent investor rule, which assumes expertise in investing, and is a higher standard than the more general reasonable person standard used in general tort cases. This is contextually informed level of responsibility becomes a field-specific subsidiary duty of care.

AI Safety researchers have identified many potential AI errors that could cause accidental harms[6], including:

- negative side effects, when an agent disturbs its environment in negative ways not accounted for in its reward function [8, 71]
- reward hacking, when an agent finds a way to optimize its reward in an unexpected and unintended way [48, 142]
- distributional shift, when the agent is trained in an environment which does not generalize to the environment in which it operates [68, 73]

A full discussion of these errors is beyond the scope of this paper. Here, we note that the fiduciary duty of care is a legal requirement that may include remediation of any or all of these problems as the context-appropriate standard of prudence evolves to take them into account.

For Fiduciary AI, the duty of care may be violated if the designer's judgment was in bad faith or not prudently informed of the facts.[24] For example, a duty of care for Fiduciary AI might require that the AI designer has considered all potential harms caused by their system. For example, consider what happens if a hypothetical social media service, *Zmeta*, is tried in court for violation of fiduciary duties after it is found to be addictive and harmful to its users. The company's lawyers argue that engagement is a signal of user interest, and so the company was acting in good faith. A court might consider a *prudent user rule* informed by research about addiction and well-being (such as [75, 102, 125]). Under such a rule, using engagement as a proxy for user best interest would be considered negligent, because it falls below the standard of care appropriate for the context.

Several reporting or labeling frameworks have been proposed for guiding due diligence and accountability for machine learning based on the data set they use for training [45], conformity with testing standards [10], use of the model only for its intended purposes and with testing for fairness on varying demographic groups [87], and for overall design of the reward system [48]. This can be framed as an improvement to AI Safety due to the reduction of epistemic uncertainty [89, 132, 134], which is uncertainty about which what can be known in principle but has not been discovered in practice. These standards are likely to evolve over time as AI, and society's understanding of it, matures.

9.1 Inductive bias

While Fiduciary AI duties apply to everything about a system, from design to deployment, some aspects of it are especially prone to negligence. Whereas the choice of training data and benchmarking of a model are often deliberate choices, the inductive bias of a machine learning system can be an afterthought, determined by defaults. This makes attention to inductive bias a good example of what is addressed by the duty of care.

We illustrate this simply in Bayesian terms. For a particular application of the system that uses the model trained on a data set D to determine the outcome for a particular case h. For simplicity, let h be a binary decision, such as whether or not to hire an individual. The closer that the likelihood ratio $\frac{P(D|h=1)}{P(D|h=0}$ is to 1, the less the available training data has informed the decision, and the more the inductive bias $P(h)$ dominates. This can happen when there is insufficient data in the region that is informative for the case of h, such as if the person h is a rare minority among a larger population. A duty of care for Fiduciary AI would standardize the inductive bias of systems based on their context and purpose of use. Notably, choices about inductive bias must be made if reward modeling is used to assess the best interests of the principal (see Section 6.1.2). Xu et al. [141] show how training on other tasks can be used to inform the prior used in IRL on a new task. Hence, a duty of care might standardize a prior over rewards directly, but it might also standardize a corpus of analogous tasks used to train that prior.

[23] Aguirre et al. [5] argue that, even in absence of stronger fiduciary duties, for AI the presence of conflicts in design must be "transparently and saliently indicate[d] to users", but not forbidden outright. This amounts to extending the duty of disclosure to all AI systems with users.

[24] In analogous duty of care cases, corporate directors are only held to the standard of making the decision with a good faith business judgement, even if their judgement was (ex post) incorrect. The burden is on the plaintiff to show bad faith or impropriety on the part of the director. This is called the *business judgement rule.*

10 DISCUSSION

Lawmakers have struggled to find a way to regulate artificial intelligence directly. At the heart of the problem is the incompleteness with which principals can come to agreement with an AI system, whether through explicit direction or consent [56]. This means that for the foreseeable future, AI will only be as ethical as the purposes of the social actor that operates it [16]. Fiduciary duties are a time-tested legal means for establishing the trustworthiness of an incompletely contracted agent by aligning its purposes with a principal [35]. Fiduciary duties for computational systems are part of the law today, and may become more broadly applied with new legislation in the future [13, 105]

This article has aimed to inform system designers about the technical requirements implied by fiduciary duties. To this end, we have outlined the legal rationale for fiduciary duties, and in particular their application to computational systems and AI. We have then provided a guide for how a Fiduciary AI system can be designed or audited for compliance with these rules. This six-step process involves: understanding the context of the system; identifying the principals; assessing the best interests of the principals; aggregating those best interests; ensuring the system is loyal to those best interests; and observing a contextually appropriate standard of care with respect to unlearned aspects of the system, such as inductive bias.

We have identified dimensions of Trustworthy AI that inform best practices in each of these stages. Notably, we see Fiduciary AI practice as drawing on both Contextual Integrity [92] and AI Alignment [143] research in an original way, while also incorporating other dimensions of Trustworthy AI. Though Fiduciary AI functions may one day be performed by a general AI system, its duties are scoped to a particular legal context in order to prevent conflicts of interest with the principals and risk more generally. The legal structure of fiduciary duties – abstract primary duties and more contextually specific subsidiary duties – is an important feature of Fiduciary AI, as in many cases legal constraints can make complex AI design challenges more tractable or straightforward.

There are many open research questions raised by the application of fiduciary principles to computational systems. We leave as an open question for future work how human-centered and participatory design methods [19, 65, 83, 116] might, or might not, be used to assess the best interests of the principals of an information fiduciary. We also wonder how the narrowly legal duty of care discussed in this paper compares with other notions of care in AI [69] and digital infrastructure [126] design. Our focus on AI and legal methods, to the exclusion of more humanistic methods, is a limitation of this article.

Fiduciary AI standards will necessarily evolve with the maturation of policy, technology, and the understanding of the courts. It is pragmatic to take stock of the state of the art in legal and technical research at this stage so that different research communities can find common ground in Fiduciary AI research, and demonstrate the feasibility of policy positions that anticipate future iterations of AI research and practice. Such policies may have far-reaching implications for value alignment of more powerful AI to come [5, 117].

ACKNOWLEDGMENTS

We gratefully acknowledge the helpful comments of Andrew Critch, Francesca Episcopo, Katrina Geddes, Jiaying Jiang, Tomer Kenneth, Aniket Kesari, Galen Panger, Ira Rubenstein, Katherine Strandburg, Thomas Streinz, Emily Tseng, Kush Varshney, Mark Verstraete, Stav Zeitouni. We also are grateful for the feedback from participants in the 3rd Annual Symposium for Applications of Contextual Integrity, Center for Human-Compatible Artificial Intelligence Beneficial AI Seminar, Northeast Privacy Scholars Workshop, and the Information Law Institute and Privacy Research Group at New York University School of Law. This material is based upon work supported by the National Science Foundation under Grant No. 2105301. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors. One of the authors of this article was supported by the New York University Information Law Institute's Fellows program, which is funded in part by Microsoft Corporation.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*. 1.

[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 252–260.

[3] Mark Ackerman, Trevor Darrell, and Daniel J Weitzner. 2001. Privacy in context. *Human–Computer Interaction* 16, 2-4 (2001), 167–176.

[4] Alessandro Acquisti and Jens Grossklags. 2003. Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior. In *2nd Annual Workshop on Economics and Information Security-WEIS*, Vol. 3. Citeseer, 1–27.

[5] Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B Reiner. 2020. AI loyalty: a new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society* 1, 3 (2020), 128–137.

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[7] Stuart Armstrong, Jan Leike, Laurent Orseau, and Shane Legg. 2020. Pitfalls of learning a reward function online. *arXiv preprint arXiv:2004.13654* (2020).

[8] Stuart Armstrong and Benjamin Levinstein. 2017. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720* (2017).

[9] Stuart Armstrong and Sören Mindermann. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in Neural Information Processing Systems* 31 (2018).

[10] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[11] Gustaf Arrhenius. 2000. An impossibility theorem for welfarist axiologies. *Economics & Philosophy* 16, 2 (2000), 247–266.

[12] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. 2018. Learning to understand goal specifications by modelling reward. *arXiv preprint arXiv:1806.01946* (2018).

[13] Jack M Balkin. 2015. Information fiduciaries and the first amendment. *UCD Law Review* 49 (2015), 1183.

[14] Jack M Balkin. 2020. The Fiduciary Model of Privacy. *Harvard Law Review Forum* 134 (2020), 11.

[15] Seth D Baum. 2020. Social choice ethics in artificial intelligence. *AI & Society* 35, 1 (2020), 165–176.

[16] Sebastian Benthall and Jake Goldenfein. 2021. Artificial Intelligence and the Purpose of Social Systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI Ethics and Society (AIES'21)*.

[17] Sebastian Benthall, Seda Gürses, and Helen Nissenbaum. 2017. *Contextual Integrity Through the Lens of Computer Science*. Now Publishers.

[18] Uri Benzion, Amnon Rapoport, and Joseph Yagil. 1989. Discount rates inferred from decisions: An experimental study. *Management Science* 35, 3 (1989), 270–284.

[19] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[20] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. " I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!"-Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021*. 763–776.

[21] Richard RW Brooks. 2019. The Economics of Fiduciary Law. In *The Oxford Handbook of Fiduciary Law*.

[22] Christopher Burr, Mariarosaria Taddeo, and Luciano Floridi. 2020. The ethics of digital well-being: a thematic review. *Science and Engineering Ethics* (2020), 1–31.

[23] Ryan Calo. 2013. Digital market manipulation. *George Washington Law Review* 82 (2013), 995.

[24] Christopher Carroll, Jiri Slacalek, Kiichi Tokuoka, and Matthew N White. 2017. The distribution of wealth and the marginal propensity to consume. *Quantitative Economics* 8, 3 (2017), 977–1020.

[25] Christopher F Chabris, David Laibson, Carrie L Morris, Jonathon P Schuldt, and Dmitry Taubinsky. 2008. Individual laboratory-measured discount rates predict field behavior. *Journal of Risk and Uncertainty* 37, 2 (2008), 237–269.

[26] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[27] Julie E Cohen. 2016. The regulatory state in the information age. *Theoretical Inquiries in Law* 17, 2 (2016), 369–414.

[28] IEEE Standards Committee et al. 2020. *IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being: IEEE Standard 7010-2020*. IEEE.

[29] Michael A Cusumano, Annabelle Gawer, and David B Yoffie. 2019. *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*. Harper Business New York.

[30] Anupam Datta, Jeremiah Blocki, Nicolas Christin, Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Arunesh Sinha. 2011. Understanding and protecting privacy: Formal semantics and principled audit mechanisms. In *International Conference on Information Systems Security*. Springer, 1–27.

[31] Daniel Dewey. 2011. Learning what to value. In *International Conference on Artificial General Intelligence*. Springer, 309–314.

[32] Anind K Dey, Gregory D Abowd, and Daniel Salber. 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction* 16, 2-4 (2001), 97–166.

[33] Ed Diener, Ed Sandvik, and William Pavot. 2009. Happiness is the frequency, not the intensity, of positive versus negative affect. In *Assessing Well-being*. Springer, 213–231.

[34] Paul Dourish. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8, 1 (2004), 19–30.

[35] Frank H Easterbrook and Daniel R Fischel. 1993. Contract and fiduciary duty. *The Journal of Law and Economics* 36, 1, Part 2 (1993), 425–446.

[36] Peter Eckersley. 2018. Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *arXiv preprint arXiv:1901.00064* (2018).

[37] Serge Egelman and Eyal Peer. 2015. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*. 16–28.

[38] IAB Europe. 2020. IAB Europe Transparency & Consent Framework Policies. https://iabeurope.eu/wp-content/uploads/2020/11/TCFv2-0Policyversion2020-11-18-3.2a.docx-1.pdf

[39] Tom Everitt, Ryan Carey, Eric Langlois, Pedro A Ortega, and Shane Legg. 2021. Agent incentives: A causal perspective. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence,(AAAI-21). Virtual. Forthcoming*.

[40] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese* 198, 27 (2021), 6435–6467.

[41] Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. 2020. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540* (2020).

[42] Tamar Frankel. 2011. Fiduciary law in the twenty-first century. *Boston University Law Review* 91 (2011), 1289.

[43] Batya Friedman, Edward Felten, and Lynette I Millett. 2000. Informed consent online: A conceptual model and design principles. *University of Washington Computer Science & Engineering Technical Report 00–12–2* 8 (2000).

[44] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. 2019. From language to goals: Inverse reinforcement learning for vision-based instruction following. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*.

[45] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[46] Martin Gelter and Geneviève Helleringer. 2019. Fiduciary Principles in European Civil Systems. In *The Oxford Handbook of Fiduciary Law*.

[47] Allan Gibbard. 1973. Manipulation of voting schemes: a general result. *Econometrica: Journal of the Econometric Society* (1973), 587–601.

[48] Thomas Krendl Gilbert, Sarah Dean, Tom Zick, and Nathan Lambert. 2022. Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems. *arXiv preprint arXiv:2202.05716* (2022).

[49] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. 2020. Quantifying differences in reward functions. *arXiv preprint arXiv:2006.13900* (2020).

[50] Andrew Gold. 2019. The Fiduciary Duty of Loyalty. In *The Oxford Handbook of Fiduciary Law*.

[51] Carol Graham, Kate Laffan, and Sergio Pinto. 2018. Well-being in metrics and policy. *Science* 362, 6412 (2018), 287–288.

[52] Leonard Green and Joel Myerson. 2004. A discounting framework for choice with delayed and probabilistic rewards. *Psychological bulletin* 130, 5 (2004), 769.

[53] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in Neural Information Processing Systems* 26 (2013).

[54] James Grimmelmann. 2019. When All You Have Is A Fiduciary. https://lpeproject.org/blog/when-all-you-have-is-a-fiduciary/

[55] Nyoman Gunantara. 2018. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering* 5, 1 (2018), 1502242.

[56] Dylan Hadfield-Menell and Gillian K Hadfield. 2019. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 417–422.

[57] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. 2017. Inverse reward design. *Advances in Neural Information Processing Systems* 30 (2017).

[58] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems* 29 (2016), 3909–3917.

[59] Woodrow Hartzog and Neil M. Richards. 2022. Legislating Data Loyalty. *Notre Dame Law Review Reflection* 97 (2022), 365.

[60] Maximilian Hils, Daniel W Woods, and Rainer Böhme. 2020. Measuring the emergence of consent management on the web. In *Proceedings of the ACM Internet Measurement Conference*. 317–332.

[61] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in Atari. *Advances in Neural Information Processing Systems* 31 (2018).

[62] Michael Jensen and William Meckling. 1976. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics* 3 (1976), 305–360.

[63] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems* 33 (2020), 4415–4426.

[64] Daniel Kahneman and Alan B Krueger. 2006. Developments in the measurement of subjective well-being. *Journal of Economic Perspectives* 20, 1 (2006), 3–24.

[65] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.

[66] Daniel B. Kelly. 2019. Fiduciary Principles in Fact-Based Fiduciary Relationships. In *The Oxford Handbook of Fiduciary Law*. Oxford Handbooks.

[67] Lina M Khan and David E Pozen. 2019. A skeptical view of information fiduciaries. *Harvard Law Review* 133 (2019), 497.

[68] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. 2020. Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2174–2184.

[69] Bran Knowles, Jasmine Fledderjohann, John T Richards, and Kush R Varshney. 2023. Trustworthy AI and the Logics of Intersectional Resistance. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 172–182.

[70] W Bradley Knox and Peter Stone. 2010. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. Citeseer, 5–12.

[71] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. 2018. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186* (2018).

[72] Ethan Kross, Philippe Verduyn, Emre Demiralp, Jiyoung Park, David Seungjae Lee, Natalie Lin, Holly Shablack, John Jonides, and Oscar Ybarra. 2013. Facebook use predicts declines in subjective well-being in young adults. *PloS One* 8, 8 (2013), e69841.

[73] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153* (2020).

[74] Per Krusell and Anthony A Smith, Jr. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy* 106, 5 (1998), 867–896.

[75] Jungwon Kuem, Soumya Ray, Pei-Fang Hsu, and Lara Khansa. 2020. Smartphone addiction and conflict: An incentive-sensitisation perspective of addiction for Information Systems. *European Journal of Information Systems* (2020), 1–22.

[76] David Laibson. 1997. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics* 112, 2 (1997), 443–478.

[77] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871* (2018).

[78] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI safety gridworlds. *arXiv preprint arXiv:1711.09883* (2017).

[79] Myles-Jay Linton, Paul Dieppe, and Antonieta Medina-Lara. 2016. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. *BMJ open* 6, 7 (2016), e010641.

[80] Zengrui Liu, Umar Iqbal, and Nitesh Saxena. 2022. Opted Out, Yet Tracked: Are Regulations Enough to Protect Your Privacy? https://doi.org/10.48550/ARXIV.2202.00885

[81] Harry M Markowitz. 1968. *Portfolio selection.* Yale University Press.

[82] R Timothy Marler and Jasbir S Arora. 2004. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization* 26, 6 (2004), 369–395.

[83] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).

[84] Alice E Marwick et al. 2018. Privacy at the margins| understanding privacy at the margins—introduction. *International Journal of Communication* 12 (2018), 9.

[85] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *ISJLP* 4 (2008), 543.

[86] Paul B. Miller. 2019. The Identification of Fiduciary Relationships. In *The Oxford Handbook of Fiduciary Law.* Oxford Handbooks.

[87] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229.

[88] Brent Mittelstadt. 2017. From individual to group privacy in big data analytics. *Philosophy & Technology* 30, 4 (2017), 475–494.

[89] Niklas Möller. 2012. The concepts of risk and safety. *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk* 1 (2012), 55–85.

[90] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *ICML*, Vol. 1. 2.

[91] Shachar Nir. 2020. One duty to all: the fiduciary duty of impartiality and stockholders' conflict of interest. *Hastings Business Law Journal* 16 (2020), 1.

[92] Helen Nissenbaum. 2020. *Privacy in context.* Stanford University Press.

[93] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. 2019. Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development* 63, 4/5 (2019), 2–1.

[94] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[95] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.

[96] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[97] Ted O'Donoghue and Matthew Rabin. 2001. Choice and procrastination. *The Quarterly Journal of Economics* 116, 1 (2001), 121–160.

[98] Andrzej Osyczka. 1984. *Multicriterion optimisation in engineering.* Halsted Press.

[99] Galen Panger. 2018. People tend to wind down, not up, when they browse social media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.

[100] Derek Parfit. 1984. *Reasons and persons.* OUP Oxford.

[101] Alexander Peysakhovich. 2019. Reinforcement learning and inverse reinforcement learning with system 1 and system 2. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 409–415.

[102] Greta L Polites, Christina Serrano, Jason Bennett Thatcher, and Kevin Matthews. 2018. Understanding social networking site (SNS) identity from a dual systems perspective: an investigation of the dark side of SNS use. *European Journal of Information Systems* 27, 5 (2018), 600–621.

[103] Mahendra Prasad. 2018. Social choice and the value alignment problem. *Artificial intelligence safety and security* (2018), 291–314.

[104] Julia M Puaschunder. 2021. Data Fiduciary in Order to Alleviate Principal–Agent Problems in the Artificial Big Data Age. In *Information for Efficient Decision Making: Big Data, Blockchain and Relevance.* World Scientific, 41–90.

[105] Neil M Richards and Woodrow Hartzog. 2020. A Duty of Loyalty for Privacy Law. *Available at SSRN* (2020).

[106] Sasha Romanosky, Alessandro Acquisti, Jason Hong, Lorrie Faith Cranor, and Batya Friedman. 2006. Privacy patterns for online interactions. In *Proceedings of the 2006 Conference on Pattern Languages of Programs.* 1–9.

[107] Constantin A Rothkopf and Christos Dimitrakakis. 2011. Preference elicitation and inverse reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 34–48.

[108] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Penguin.

[109] Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, and Vincent Roca. 2021. Consent Management Platforms under the GDPR: processors and/or controllers?. In *Annual Privacy Forum.* Springer, 47–69.

[110] Steven L Schwarcz. 2009. Fiduciaries with Conflicting Obligations. *Minn. L. Rev.* 94 (2009), 1867.

[111] Paul M Schwartz and Karl-Nikolaus Peifer. 2017. Transatlantic data privacy law. *Georgia Law Journal* 106 (2017), 115.

[112] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 59–68.

[113] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. 2019. Preferences implicit in the state of the world. *arXiv preprint arXiv:1902.04198* (2019).

[114] Robert H Sitkoff. [n. d.]. Other Fiduciary Duties. In *The Oxford Handbook of Fiduciary Law.*

[115] Robert H Sitkoff. 2011. The economic structure of fiduciary law. *Boston University Law Review* 91 (2011), 1039.

[116] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–6.

[117] Nate Soares and Benja Fallenstein. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report* 8 (2014).

[118] Daniel J Solove. 2021. The myth of the privacy paradox. *George Washington Law Review* 89 (2021), 1.

[119] Shawn Squire, Stefanie Tellex, Dilip Arumugam, and Lei Yang. 2015. Grounding English commands to reward functions. In *Robotics: Science and Systems.*

[120] Wolfram Stadler. 1988. Fundamentals of multicriteria optimization. In *Multicriteria Optimization in Engineering and in the Sciences.* Springer, 1–25.

[121] Thomas Streinz. 2021. The Evolution of European Data Law. *The Evolution of EU Law (OUP, 3rd edn 2021)* (2021), 902–936.

[122] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet Policy Review* 8, 2 (2019).

[123] Linnet Taylor, Luciano Floridi, and Bart Van der Sloot. 2016. *Group Privacy: New Challenges of Data Technologies.* Vol. 126. Springer.

[124] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[125] Morten Tromholt. 2016. The Facebook experiment: Quitting Facebook leads to higher levels of well-being. *Cyberpsychology, Behavior, and Social Networking* 19, 11 (2016), 661–666.

[126] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. 2022. Care infrastructures for digital security in intimate partner violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–20.

[127] Andrew Tuch. Forthcoming 2021. A General Defense of Information Fiduciaries. *Washington University Law Review* 98 (Forthcoming 2021).

[128] Amos Tversky and Daniel Kahneman. 1985. The framing of decisions and the psychology of choice. In *Behavioral Decision Making.* Springer, 25–41.

[129] Amos Tversky and Daniel Kahneman. 1989. Rational choice and the framing of decisions. In *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers.* Springer, 81–126.

[130] Werner Ulrich. 1987. Critical heuristics of social systems design. *European Journal of Operational Research* 31, 3 (1987), 276–283.

[131] Jennifer M Urban and Chris Jay Hoofnagle. 2014. The privacy pragmatic as privacy vulnerable. In *Symposium on Usable Privacy and Security (SOUPS 2014) Workshop on Privacy Personas and Segmentation (PPS).*

[132] Kush R Varshney. 2016. Engineering safety in machine learning. In *2016 Information Theory and Applications Workshop (ITA).* IEEE, 1–5.

[133] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.

[134] Kush R Varshney and Homa Alemzadeh. 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data* 5, 3 (2017), 246–255.

[135] Salome Viljoen. 2021. A relational theory of data governance. *Yale Law Journal* 131 (2021), 573.

[136] Salomé Viljoen, Jake Goldenfein, and Lee McGuigan. 2021. Design choices: Mechanism design and platform capitalism. *Big Data & Society* 8, 2 (2021), 20539517211034312.

[137] Ari Ezra Waldman. 2019. Privacy Law's False Promise. *Washington University Law Review* 97 (2019), 773.

[138] Ari Ezra Waldman. 2021. *Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power.* Cambridge University Press.

[139] Michael Walzer. 2008. *Spheres of Justice: A Defense of Pluralism and Equality.* Basic books.

[140] Jeannette M Wing. 2021. Trustworthy AI. *Commun. ACM* 64, 10 (2021), 64–71.

[141] Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. 2019. Learning a prior over intent via meta-inverse reinforcement learning. In *International Conference on Machine Learning*. PMLR, 6952–6962.

[142] Yinlong Yuan, Zhu Liang Yu, Zhenghui Gu, Xiaoyan Deng, and Yuanqing Li. 2019. A novel multi-step reinforcement learning method for solving reward hacking. *Applied Intelligence* 49, 8 (2019), 2874–2888.

[143] Eliezer Yudkowsky. 2016. The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker* (2016).

[144] Ewa Zawojska and Mikołaj Czajkowski. 2017. Re-examining empirical evidence on stated preferences: importance of incentive compatibility. *Journal of Environmental Economics and Policy* 6, 4 (2017), 374–403.

[145] Simon Zhuang and Dylan Hadfield-Menell. 2020. Consequences of misaligned AI. *Advances in Neural Information Processing Systems* 33 (2020), 15763–15773.

[146] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *AAAI*, Vol. 8. Chicago, IL, USA, 1433–1438.