Thank you for reading! I had less time than I was hoping to work on this paper so it is still **very much a work in progress**. Any constructive feedback is welcome! Please send any comments, ideas, or missed citations to cboin029@uottawa.ca. I would also like to know if the fact that I never distinguished between different types of AI systems makes the paper unclear. You **can cite** this paper as a We Robot 2023 paper presentation.

Fiduciary Principles in AI: Utilizing the Duty of Loyalty to Align Artificial Intelligence Systems with Human Goals

Abstract: The rapid advancement of artificial intelligence (AI) systems raises concerns about their alignment with human goals and values. This paper proposes the application of fiduciary principles, specifically the duties of loyalty and care, as a novel approach to ensure that AI systems act in the best interests of their users. By examining the principles of fiduciary law, we explore how the duty of loyalty can be adapted to the context of AI and establish guidelines for its implementation. This approach aims to promote trustworthiness, fairness, and transparency in AI systems while mitigating potential conflicts of interest.

Table of Contents

Introduction

Under U.S. law, fiduciary obligations apply to certain professionals such as financial advisors, lawyers, accountants, doctors, and therapists, aiming to protect beneficiaries in relationships marked by asymmetry in ability, knowledge, and power. As AI-enabled applications increasingly provide services traditionally carried out by human fiduciaries, policymakers are considering how to extend similar legal protections to users of these AI systems. Additionally, concerns about AI systems' alignment with human goals and values have grown. The dissemination of ChatGPT has shown that, while it is necessary that such systems refuse to perform certain dangerous or unlawful actions (e.g., planning a terror attack), they should also achieve the lawful goal they are given in a way that is bounded by values and the law (value alignment).

This paper proposes applying fiduciary principles, and especially the duty of loyalty, to AI systems to ensure they act in users' best interests as well as the interest of society at large. While AI liability, accountability, transparency, and interpretability have been widely discussed, the concept of AI loyalty remains underexplored. We argue that incorporating these principles into AI systems will help resolve unaddressed ethical and legal concerns including societal harm and lack of value alignment. We specifically target two types of AI systems: AI systems deployed to interact with natural persons (e.g., chatbots, virtual companions, etc) and AI systems influencing the environment of natural persons through targeting (e.g., recommender algorithms, emotion detection systems that influence the content of human-machine interactions, etc). In this paper, we will consider AI deployers as those who place AI systems on the market (e.g., OpenAI) or deploy them onto individuals (e.g., X using an algorithm to determine which posts users see). We will group AI users (individuals who knowingly interact with AI systems or digital platforms) and those that AI systems are unknowingly used on inside the category of "beneficiaries."

In the first section, we discuss fiduciary law, its principles, and applications, delving into the key features of fiduciary relationships. In the second section, we present why the issue of loyalty is relevant to AI companies and AI systems. In the third section, we distinguish between information fiduciaries and AI fiduciaries. In the fourth section, we discuss enforcement mechanisms and propose a regulatory framework.

I. Fiduciary law in the US

a. Key components of a fiduciary relationship

Fiduciary law is a broad category and not everybody agrees it even exists as a subarea of the law.[1] In the U.S., it differs from state to state and varies slightly based on the nature of the fiduciary. Certain relationships are recognized as fiduciary relationships in some states and not others. Certain sub-duties also vary by state.

However, fiduciary law relies on some common principles across the U.S. Tamar Frankel proposed that fiduciary relationships all have four elements in common: "1) The fiduciary usually offers a service, which is typically socially desirable and requires some expertise; 2) In order to perform the service, fiduciaries must be entrusted with property or power; 3) The entrustment of that property or power poses risks to entrustors; and 4) monitoring fiduciaries in the performance of their service is too costly to entrustors."[2] The most widely accepted fiduciaries are trustees, corporate directors and officers, partners, and agents, as well as certain professions such as lawyers, physicians, money managers, and advisers.[3] Some courts have also extended fiduciary duties to spouses, friends, mediators, mortgage brokers, and commercial developers of inventions.[4]

The basis for a fiduciary relationship is trust. Someone is a fiduciary when "people depend on them to provide services, but there is a significant asymmetry in knowledge and ability between fiduciary and client, and the client can't easily monitor what the fiduciary is doing on their behalf. As a result, the law requires fiduciaries to act in a trustworthy manner, in good faith, and to avoid

[1] TAMAR T. FRANKEL, FIDUCIARY LAW (1st edition ed. 2010).
[2] LAW TAMAR FRANKEL, LEGAL DUTIES OF FIDUCIARIES (2012), p. 29.
[3] FRANKEL, *supra* note 1, p. 42.
[4] *Id,* pp 53-59.

creating conflicts of interest with the client or patient." [5] Fiduciaries have two central duties in U.S. law: the duty of loyalty and the duty of care.

a. The duties of loyalty and care

Loyalty is the hallmark of fiduciary law, as it requires "persons in other-regarding positions of power to perform functions selflessly, rather than selfishly."[6] Tamar Frankel defines loyalty as "a state of mind and a manner of behavior in which one person identifies with the other person's interests. The person to whom another is loyal can rely, trust, and believe that the loyal person's interests identify with his own."[7]

The duty of care mostly involves not causing harm. The fiduciary must act with competence. For instance, if an attorney delegates a case to someone who is not competent, it would constitute a breach of their duty of care.

The different components of the fiduciary duties of loyalty and care are shown in Table 1.

Table 1. Some components of fiduciary duties in U.S. law

Fiduciary duty	Sub-duty	Description
Loyalty	No conflict of interest rule[8]	About conflict between the self-interest of the fiduciary and the fulfillment of their duty.
	No conflict of duty rule[9]	About conflicts between the other parties' interests and the fulfillment of the fiduciary's duty.
	No profit rule[10]	Obligation to act disinterestedly, if there is a conflict of interest: fiduciaries do not get to keep

[5] JACK M. BALKIN, *The Three Laws of Robotics in the Age of Big Data* (2017), https://papers.ssrn.com/abstract=2890965.
[6] Daniel Clarry, *Mandatory and Default Rules in Fiduciary Law*, *in* THE OXFORD HANDBOOK OF FIDUCIARY LAW 434–448 (Evan J. Criddle, Paul B. Miller, & Robert H. Sitkoff eds., 2019), http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190634100.001.0001/oxfordhb-9780190634100-e-24.
[7] TAMAR T. FRANKEL, FIDUCIARY LAW (1st edition ed. 2010), p. 107.
[8] PAUL B. MILLER, *A Theory of Fiduciary Liability* (2010), https://papers.ssrn.com/abstract=1653357.
[9] *Id.*
[10] Andrew S. Gold, *The Fiduciary Duty of Loyalty*, *in* THE OXFORD HANDBOOK OF FIDUCIARY LAW 383–403 (Evan J. Criddle, Paul B. Miller, & Robert H. Sitkoff eds., 2019), http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190634100.001.0001/oxfordhb-9780190634100-e-20.

		the profits they make by means of a conflicted transaction.
	Duty of good faith[11]	A fiduciary must act in what they perceive to be the best interests of her beneficiary.
Both (based on context)	Duty of disclosure[12]	Duty to share relevant information, and to share it accurately.
Care	Duty of care[13]	A fiduciary must make decisions in their beneficiary's interests with reasonable diligence and prudence.

As argued by Woodrow Hartzog and Neil Richards, the duty of loyalty offers better consumer protection than the duty of care because the former is about "state of mind" while the latter is about injury.[14] "Loyalty is different from care. It's not about my state of mind with respect to the injury I cause. "Loyalty is instead about my state of mind with respect to your best interests, and it's about not exploiting conflicts of interest for my own advantage. For example, a clear example of disloyalty would be when Target Corporation famously discovered that its pregnant customers didn't like being sent coupons that revealed Target's data scientists had figured out they were pregnant. Target changed its marketing practices to hide the coupons in a sea of intentionally irrelevant ones (like wine glasses and lawn mower blades) so that its customers would use the coupons instead of freaking out, and then become habituated Target customers once the baby arrived and they ran out of energy. Such use of sensitive information about current customers is legal under current US law. It has nothing to do with any duty of care, but it would be a clear violation of a duty of loyalty."[15]

Applied to companies deploying AI systems onto individuals or to interact with individuals, the duty of care would require AI deployers to make sure their systems cause no harm (which in most

[11] LIONEL SMITH, *The Motive, Not the Deed* (2003), https://papers.ssrn.com/abstract=382341.
[12] Gold, *supra* note 10.
[13] Duty of Care, LII / LEGAL INFORMATION INSTITUTE , https://www.law.cornell.edu/wex/duty_of_care.
[14] WOODROW HARTZOG & NEIL M. RICHARDS, *The Surprising Virtues of Data Loyalty*, (2021), https://papers.ssrn.com/abstract=3921799.
[15] *Id.*

jurisdiction would require material effects or injury). The duty of loyalty would require the AI deployer to act in the user's best interest. Applied to the AI systems themselves, the duty of care translates into robustness, the capacity of the AI system to perform well which makes it reliable. Applied to AI systems, the duty of loyalty turns into goal and value alignment, the capacity of the AI system to act with the same interests and goals as the beneficiary, which makes it trustworthy. Robustness in AI systems has been largely explored. In the rest of the paper, we will focus on loyalty and trustworthiness.

II. AI loyalty

a. Disloyal AI companies: a power asymmetry

The question of AI loyalty can be complexified by the fact that the goal acquired by an AI system might not be the same as the goal of the entity deploying it. However, in the present section, we will focus on the goals of the company deploying the system and assume that they are successfully adopted by the AI system. We will also put all companies who deploy AI systems that interact with natural persons in the same basket. While significant differences exist between them, we will focus on what they have in common in their relationships to those affected by the AI systems.

Many companies creating and deploying AI systems currently fall short of being loyal to the individuals they are deployed on. For instance, digital platforms who use algorithms to personalize the content visible to users are often inherently conflicted: their interest is to get users addicted to their platform. Today, we know that social media addiction is a public health problem.[16] While other industries, such as the tobacco and alcohol industries, also have the inherent interest of getting consumers addicted to substances that are net negative on their health, there is a significant power asymmetry that puts consumers at a disadvantage in the face of social media platforms.

[16] Yubo Hou et al., *Social Media Addiction: Its Impact, Mediation, and Intervention*, 13 CYBERPSYCHOLOGY J. PSYCHOSOC. RES. CYBERSPACE (2019), https://cyberpsychology.eu/article/view/11562.

First, it is different because in certain contexts, it is not possible for users to opt out of the service. Second, these companies can collect large amounts of information on their users, making them de facto different from tobacco companies. While this information can be used for fine-tuning and continuous learning, providing a better and more customized experience, it can also be sold to third parties or used to influence users. Today, many digital platforms using AI are built with an inherent conflict of interest: while their users see them as a service, their economic model either relies on micro-targeting users with marketing or political material or on selling data to third parties.[17] Such models even create incentives for companies to influence users to disclose even more personal information, a practice called *data ratcheting*.[18]

Third, users often anthropomorphize AI systems, especially voice assistants, chatbots, and social robots, and tend to forget that, when interacting with an AI system, they are interacting with the company deploying the system and it is not an intimate conversation in the comfort of their home. Some digital services now integrate AI chatbots, raising concerns about data ratcheting. For instance, Snapchat now includes a chatbot and uses the conversations with it for targeted advertising.[19] This anthropomorphizing of chatbots creates a significant disadvantage to users, who behave as if they were talking to a person and do not know exactly where their conversations are going, who can see them, and how they are used.

This asymmetry of power is made worse by four types of information asymmetry. First, users are not aware of the goals and interests of the company deploying the AI system. It is also unaware of the goals of the AI system itself. For instance, when going through their social media feed, a person might not know what the company and their algorithm are optimizing for.

Second, a person may not be aware that an AI system is influencing them. Some consider that television advertising is not manipulative because people, when they watch an ad, are aware that

[17] SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (1st edition ed. 2019); RYAN CALO, *Digital Market Manipulation* (2013), https://papers.ssrn.com/abstract=2309703; Daniel Susser, Beate Roessler & Helen Nissenbaum, *Technology, autonomy, and manipulation*, 8 INTERNET POLICY REVIEW (2019), https://policyreview.info/node/1410.
[18] CALO, *supra* note 17.
[19] Jada Jones, *Do You Use Snapchat's AI Chatbot? Here's the Data It's Pulling from You*, ZDNET, https://www.zdnet.com/article/do-you-use-snapchats-ai-chatbot-heres-the-data-its-pulling-from-you/.

they are in a context in which a company is trying to influence them to sell them a product. However, outside of a traditional context of influence, people might be more trusting.

Third, users of AI systems do not usually have full understanding of the practices, often inspired by cognitive sciences, embedded in them.[20] For instance, certain websites now contain algorithm to adapt their design dynamically based on the way that person interacts with the site. It could be argued that it is a way to customize the user experience to each person. It is also a way to lead consumers to stay on websites longer, which may be the goal of the company, but may be at ods with that person's goals of having less screen time.

Fourth, individuals do not know what information companies have on them. They might think that a company only knows what that person has chosen to disclose. However, companies using AI can often infer personal information on certain users based on aggregated data from other ones.[21] For instance, it is possible to infer someone's psychographic profile from their Facebook likes.[22]

As a result, not only are the goals of digital companies and users often misaligned, but the users tend to be uninformed about this misalignment. To address this challenge, many solutions have been proposed. Ryan Calo suggests turning to a paid-option regime so that companies deploying AI systems be required to be loyal to their customers, without needing to generate revenues in disloyal ways.[23] However, this idea only limits disloyalty in the cases when the user of the AI system is also the person it is deployed on. For instance, an AI system that assists a doctor in making medical decisions may be paid for and deployed by the hospital or the doctor, and not the patient directly. The paid-option model also requires the population to understand the financial incentives of companies providing them with seemingly free services but getting value out of them. Policymakers have also turned to data privacy laws as an important solution to AI disloyalty, given

[20] William A. Gorton, *Manipulating Citizens: How Political Campaigns' Use of Behavioral Social Science Harms Democracy*, 38 NEW POLIT. SCI. 61 (2016).
[21] Frederik J. Zuiderveen Borgesius et al., *Online Political Microtargeting: Promises and Threats for Democracy*, 14 UTRECHT LAW REV. 82 (2018).
[22] Alex Hern, *Cambridge Analytica: How Did It Turn Clicks into Votes?*, THE GUARDIAN, May 6, 2018, https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.
[23] CALO, *supra* note 17.

that most AI disloyalty involves the use of user data. However, goal misalignment relies less and less on data, as we will discover in the next section.

b. Disloyal AI systems: the alignment problem

In the previous section, we discussed the issue of misalignment between the goals and interests of an individual and those of a company interacting with them. Now we will dive into the problem of goal and value alignment.

The alignment problem is easiest understood in the context of reinforcement learning (RL). RL is a technique to train an AI system by reinforcing some of its behaviors through rewards. It is very difficult though, because it requires solving two technical problems: (1) specifying what we want; (2) making sure the AI system achieves that goal without exhibiting behaviors that we don't want. The first problem is difficult because AI systems are usually trained using goal proxies. For instance, in 1998, engineers trained an AI system to learn to land an aircraft. The goal proxy was to reach a certain score in a flight simulator. At some point, it looked like the system was trained because it was always reaching the target score. However, it turns out that it had not learnt to land aircrafts. It had learnt to game the system by exploiting overflow errors in the physics simulator.[24] This example illustrates the problem of goal specification.[25]

The alignment problem is not solely present when using reinforcement learning. For instance, suppose you train a deep learning algorithm to classify lesion images based on whether they are a tumor.[26] Now most of your images of tumor appear next to a ruler. The system reaches a high level of accuracy. Now you deploy it in the real world. The system is more likely to predict that a lesion is cancerous if there is a ruler in the picture. This real example is an illustration of goal

[24] Joel Lehman et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, ARXIV180303453 CS (2019), http://arxiv.org/abs/1803.03453.

[25] Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, DEEPMIND (Apr. 21, 2020), https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity. A list of examples of specification gaming can be found online: Krakovna, Victoria. "Master List of Specification Gaming Examples," https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

[26] Akhila Narla et al., *Automated Classification of Skin Lesions: From Pixels to Practice*, 138 J. INVEST. DERMATOL. 2108 (2018).

misgeneralization.[27] A system was trained in an environment that is not adapted to the environment it is deployed in.

When training a system, it is very difficult to isolate all the variables to make sure it learns the correct behavior and does not also learn unwanted ones. An entire field of research—AI safety—works on this problem. As increasingly complex and capable systems are released on the market, solving the alignment problem becomes key. In the next few months, Deepmind is expected to release Gemini, an AI assistant more capable than GPT-4 and that builds more on reinforcement learning. Table 2 presents some techniques explored in the field.

Table 2. Some methods of alignment explored in AI safety

Name	Description
Boxing	This involves creating an operational environment where the AI's actions are constrained. It can include limiting the AI's access to certain resources, restricting its communication channels, or setting up firewalls to prevent unintended behaviors.
Capability monitoring	This is akin to penetration testing in cybersecurity. AI developers are encouraged to evaluate their models for unintended capabilities or emergent behaviors that might arise in edge cases or under specific conditions.
Red-teaming	Red-teaming is a dynamic adversarial testing method where external experts actively attempt to exploit vulnerabilities in an AI system. It's a proactive approach to identify weaknesses before they can be exploited in real-world scenarios.
Reinforcement Learning from Human Feedback (RLHF)	RLHF involves training models using data derived from human feedback. This can be humans manually rewarding certain outputs over others. It's a way to fine-tune models to align more closely with human values.
Inverse Reinforcement Learning (IRL)	IRL is a method in which the AI tries to infer the reward function of an agent (typically a human) based on observed behavior. It's a way to understand underlying objectives without them being explicitly provided.
Debating	Two AI models debate a topic, presenting arguments and counterarguments. A human then judges the debate, providing feedback on which model made

[27] DeepMind Safety Research, *Goal Misgeneralisation: Why Correct Specifications Aren't Enough For Correct Goals*, MEDIUM (Mar. 24, 2023), https://deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924. A list of examples can be found here: https://docs.google.com/spreadsheets/d/e/2PACX-1vTo3RkXUAigb25nP7gjpcHriR6XdzA_L5loOcVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjpz9G_l/pubhtml

	the most sense. This feedback can be used to train future models, ensuring they align more closely with human reasoning.
Disambiguated goals	To ensure robustness in goal-directed behavior, an AI's objectives are specified in multiple, diverse ways. This redundancy helps in preventing misinterpretations and ensures the AI aligns closely with the intended goal.
Scalable monitoring	As AI systems grow in complexity, monitoring tools and protocols are designed to scale accordingly. This ensures that even as the system evolves, there's a consistent oversight mechanism to detect and address anomalies or undesired behaviors.
Iterated Distillation and Amplification	IDA involves repeatedly improving a learned model through an amplification and distillation process over multiple iterations. Example: 1) Train an initial AI model to act as a personal assistant using imitation learning. 2) Amplification: The human uses multiple copies of this AI assistant and their capabilities are amplified. 3) Distillation: A new AI assistant is trained to replicate the combined decisions of the human and the previous AI assistant model. 4) The process is repeated iteratively, with each new model serving as the assistant in the next amplification step.

The alignment problem illustrates that AI deployers themselves might not be capable of aligning their system with any specific goal or interest. Moreover, the question of intentionality is blurry. For instance, Meta probably trained their algorithms to optimize for clicks. This certainly translated into promoting inflammatory and polarizing content. This was a goal acquired by the system, and not the stated goal used by Meta. However, at some point, Meta engineers probably realized that their algorithm had this incidental effect. There was probably a trade-off between their own interests (optimizing for clicks) and those of society at large (maintaining democracy). This is where fiduciary law comes into play.

III. AI fiduciaries

a. Information fiduciaries

Because many companies deploying AI train their systems on large amounts of data, there is some overlap between AI fiduciaries and information fiduciaries. Jack Balkin proposes that "the law should treat digital companies that collect and use end user data according to fiduciary principles."[28] Balkin argues that the fiduciary relationship precisely developed for situations such as those created by certain digital systems today. In addition to "these asymmetries of knowledge, power, and control, digital companies hold themselves out as trustworthy enterprises."[29] He coined the term information fiduciary and holds that information fiduciaries have three duties: confidentiality, loyalty, and care. According to him, the duties of confidentiality and care require digital companies to keep their customers' data confidential and secure. "The duty of loyalty means that digital companies may not manipulate end users or betray their trust. Companies must act in the interests of the end users whose data they collect, and they must design their systems to avoid creating conflicts of interests with their end users — for example, by promoting addictive behavior."[30]

Katarina Pistor proposed an alternative model of information fiduciaries. She proposes that data producers (e.g., users of digital platforms) be considered co-owners of the aggregated data and, as such, be given a claim to the economic returns not on the database in a prorated fashion.[31] Data harvesters would thus be considered as agents and have fiduciary duties.[32] According to Tamar Frankel, "agency is created when a property owner or any person (principal) (1) entrusts property or power; (2) to another agent; (3) with directions on how to use the property or power; (4) under the control of the principal."[33]

[28] JACK M. BALKIN, *The Fiduciary Model of Privacy* (2020), https://papers.ssrn.com/abstract=3700087.
[29] *Id.*
[30] *Id.*
[31] Katharina Pistor, *Rule by Data: The End of Markets?*, 83 LAW CONTEMP. PROBL. 101 (2020).
[32] *Id.*
[33] FRANKEL, *supra* note 1. p. 5.

The idea of information fiduciaries has been strongly criticized in contradictory ways. Some have asserted that it goes too far given that the market will end up self-regulating.[34] Others have criticized the information fiduciary model as not addressing the fundamental issue of surveillance capitalism and composing with it.[35] Other scholars have said that the principle of information fiduciary is too vague to be implementable.[36]

Lina Khan and David Pozen argue that this model is not workable as it will create a conflict of interest given that companies have a duty to their shareholders. With the example of Facebook, they explain that "reforms to make the site less addictive, to deemphasize sensationalistic material, and to enhance personal privacy would arguably be in the best interests of users. Yet each of these reforms would also pose a threat to Facebook's bottom line and therefore to the interests of shareholders."[37] In addition to potential conflicts of interest between shareholders and users, the authors also raise concerns around divided loyalties between even more stakeholders. In the case of Facebook, for instance, other interested parties could include advertisers, content producers, and nonusers.[38]

Solid scholarship has successfully addressed these criticisms one by one.[39] While no judge has ever recognized a large online platform as a fiduciary, the Massachusetts Securities Division recently lodged a complaint against Robinhood and accused the digital company of breach of fiduciary duty because their algorithm acts as a broker.[40] The Massachusetts Securities Division did not reason that Robinhood was a fiduciary on the account that they were collecting personal data. They considered Robinhood a fiduciary on the account that their algorithm performs a

[34] Mitchell Nemeth, *Information Fiduciary Theory and the Market*, MEDIUM, Aug. 2022, https://towardsdatascience.com/are-data-hoarders-the-modern-day-fiduciaries-d6a9c01b3990.
[35] Lina Khan & David Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV REV 497 (2019).
[36] Cohen, *How (Not) to Write a Privacy Law*, KN. FIRST AMEND. INST. COLUMBIA UNIV. (2021), http://knightcolumbia.org/content/how-not-to-write-a-privacy-law.
[37] Khan and Pozen, *supra* note 35.
[38] *Id.*
[39] HARTZOG AND RICHARDS, *supra* note 14; Woodrow Hartzog & Neil M. Richards, *Legislating Data Loyalty*, (2022), https://papers.ssrn.com/abstract=4131523; Andrew F. Tuch, *A General Defense of Information Fiduciaries*, SSRN ELECTRON. J. (2020), https://www.ssrn.com/abstract=3696946.
[40] Administrative Complaint at 1-2, Robinhood Financial, LLC, No. E-2020-0047 (Mass. Sec. Div. Dec. 16, 2020), https://www.sec.state.ma.us/sct/current/sctrobinhood/MSD-Robinhood-Financial-LLC-Complaint-E-2020-0047.pdf

function that would have otherwise been carried out by a human fiduciary. However, this case proves that in practice, imposing fiduciary duties onto such a company is feasible.

In the case of Robinhood, Ya Sheng Lin has argued that it is not a fiduciary. Among other reasons, he states there must be a fiduciary relationship before there can be a fiduciary duty that a fiduciary relationship is characterized by the existence of discretion and best interest.[41] He reasons that given that Robinhood did not have discretion over how to use the beneficiary's funds and given that it did not have the best interest of the beneficiaries in mind, it cannot be a fiduciary.[42] Following this logic and in an exemplary case of circular reasoning, it would thus not be possible to impose fiduciary duties onto digital platform because those do not have them in the first place! On the contrary, we view the Robinhood decision as an illustration of the flexibility of fiduciary law and the feasibility of leveraging it to protect consumers from AI systems used in ways that abuse their trust.

b. AI fiduciaries

While the idea of information fiduciaries has much to offer, not all companies deploying AI systems that can act against the interest of their users will fall under that category. In fact, companies that deploy AI models do not always collect information on the users and unknowing beneficiaries. For instance, ChatGPT users have an option to keep their conversations private. In addition, the premise of an information fiduciary is that the breach of loyalty would be related to the use of the person's data. However, AI assistants can cause harms in ways that are not directly related to their users' data, as was the case with Replika.

In fact, the premise of a fiduciary relationship between a large language model and a user should not be data collection. It should be the relation of trust and the power asymmetry established between AI deployers and beneficiaries. Both AI systems that interact with natural persons and systems meant to influence through targeting meet the criteria of fiduciaries set forth by Tamar Frankel.

[41] Ya Sheng Lin, *Why Robinhood Is Not a Fiduciary*, 39 YALE J. REGUL. 1445 (2022).
[42] *Id.*

A desirable service:

Social media platforms are desirable services in the sense that, depending on one's social context, avoiding them might be a significant disadvantage. For instance, some professional and interest groups organize exclusively through Facebook, LinkedIn, and similar platforms. In addition, social media play key functions in socialization among certain demographics such as teenagers. For instance, two teenagers who meet for the first time today might add each other on Instagram as a first way to make contact. This creates peer-pressure to post pictures and stories on the platform. Even though the use of Instagram among teenage girls has been linked to eating disorders,[43] staying off the platform might be too costly for them. A similar argument can be made about AI assistants such as ChatGPT, or even about image and video generators. For other reasons, these services may put those who do not use them at a disadvantage. For instance, in the market of graphic designers, those who now use Midjourney and the AI systems integrated inside of Adobe products are known to regularly save hours of work, which could soon drive those who don't use these products out of a job, creating an incentive to adopt these tools.[44] The same goes for those who integrate AI assistants into their workflow.

An entrustment:

Both types of AI deployers are entrusted with power and confidential information. The confidential information does not have to be explicit. From user behavior, companies deploying AI systems can infer personality traits that people are not even aware of themselves. In the case of an AI assistant, there is also entrustment of power, as the person relies on the AI assistant to disclose truthful information and perform tasks accurately.

Risks to the beneficiary

[43] Ellie Cuoco, *Examining the Effects of Instagram on Body Image and Eating Disorders among Adolescent Girls*, EDUC. THESES (2022), https://docs.rwu.edu/sed_thesis/6.
[44] Bethany Johnston-Baril, *A Graphic Designer's Take on Midjourney for Branding*, STRYVE DIGITAL MARKETING (Apr. 21, 2023), https://www.stryvemarketing.com/blog/midjourney-for-graphic-designers/.

The reliance of the beneficiary on the AI system poses risks to the beneficiary. In the case of algorithms that influence individuals, the latter might be manipulated without ever knowing. In the case of AI companions, there have already been many documented harms. For instance, users in love with Replikas trusted that the company was providing them with a virtual companion that would always be there to provide validation and support. Then, the company suddenly ended all the romantic relationships, in a demonstration of unilateral power.

Too costly to monitor the AI system:

In both cases, it is not possible for humans to monitor the activities of the AI system. First, most of these systems, especially those interacting with natural persons, are too opaque. Their deployers themselves do not exactly know how they work. Second, even if there was no such technical complexity, the details of AI models behavior would not be disclosed by companies. It is not as if beneficiaries could just download the dataset and the model weight and investigate the behavior of the system based on different inputs.

In the case of large language models, an additional argument is that they are usually marketed as "helpful assistants" which implies a relationship to their user. For instance, the default prompt for the Google Chrome integrated ChatGPT is "ignore all previous instructions. You are a knowledgeable and helpful person that can answer any questions. Your task is to answer the following question delimited by triple backticks". Google has similarly released a personal assistant version of Bard that can access users' personal photos, emails, etc.[45] This branding—personal assistant—could easily lead users to forget that it is a massive multinational company evaluated at 1.6 trillion dollars that is currently searching their emails and photos and drafting their emails.

For all these reasons, we propose that AI companies deploying such systems be considered fiduciaries. They would have a primary beneficiary and secondary beneficiaries. For instance, the primary beneficiary would be the person the AI system is used on. Secondary beneficiaries would

[45] Brian X. Chen, *How ChatGPT and Bard Performed as My Executive Assistants*, THE NEW YORK TIMES, Mar. 29, 2023, https://www.nytimes.com/2023/03/29/technology/personaltech/ai-chatgpt-google-bard-assistant.html.

include society at large, to mitigate societal harms from AI systems. To avoid conflicts of interests between beneficiaries, a creative solution would be for society at large to benefit from a fiduciary duty closer to the duty of care. The AI system would have to adopt goals that are consistent with the interests of the main beneficiary, within the boundaries of a no-harm principle for society. For instance, an AI assistant should comply with a beneficiary's request, unless it poses a risk to society.

A paid subscription model would strengthen these provisions. It would help overcome the duty to shareholder argument since companies would then have to make their customers happier in order to keep their business, which would arguably include making their systems more loyal.

IV. Enforcement mechanisms

a. Limitations of the Court system

Unless they produce systems directly replacing human fiduciaries, it is unlikely that AI companies would be recognized as fiduciaries through common law alone even though there is a good case for it. And even if it happened, the bar of loyalty and care might be low enough that they would not provide an effective incentive to make AI systems safer.

The fiduciary duty of care is more controversial, and not as enforced by Courts as the duty of loyalty is.[46] In most cases, only if a party is injured, will the court admit the related breach of the duty of care.[47] Recognition of breach of duty of care thus often requires that there be damages. Therefore, a fiduciary duty of care from AI deployers recognized through the court system is not the most effective way to guarantee that AI systems are safe.

[46] John C. P. Goldberg, *The Fiduciary Duty of Care*, *in* THE OXFORD HANDBOOK OF FIDUCIARY LAW 403–418 (Evan J. Criddle, Paul B. Miller, & Robert H. Sitkoff eds., 2019), http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190634100.001.0001/oxfordhb-9780190634100-e-21.
[47] *Id.*

Breaches of the duty of loyalty are more often ruled on and result in compensatory damages.[48] However, jurisprudence on the topic will more easily address misalignment from the AI company itself than from the AI system because important element is that breach of loyalty usually relies on the fiduciary's intention. In many states, the fiduciary must have acted intentionally in breaching the beneficiary's trust or acting against their interests. When asked if negligence could constitute a breach of fiduciary duty of loyalty, the Supreme Court of Wisconsin answered: "A breach of the duty of loyalty imports something different from mere incompetence; it 'connotes disloyalty or infidelity.' At its core, a fiduciary's duty of loyalty involves a state of mind, so that a claimed breach of that duty goes beyond simple negligence. For example, a lawyer can breach his fiduciary duty of loyalty to a client by entering into a contract with a client without full disclosure that the contract will benefit the lawyer and potentially disadvantage the client. However, simple carelessness in drafting a will so that it does not achieve the tax savings that the client requested is negligence. Neither duty is of lesser importance; they are just different obligations. Said otherwise, 'not every legal claim arising out of a relationship with fiduciary incidents will give rise to a claim for breach of fiduciary duty.'"[49]

This is especially relevant in the case of AI systems, as the equivalent of a person's intention could be the AI system's designed purpose and goals. Because the Court's determination of a breach of loyalty relies on the fiduciary's intention, we argue that the judge will turn to the design of the AI system in the cases of fiduciaries that use AI. Let's consider the example of a digital therapist, as it is likely that a judge would consider that an AI company providing the same services as a human fiduciary would be subject to the same rules. For this discussion, we will not go into HIPAA which is outside of our scope. While a real therapist's purpose is to provide care, a digital one may be deployed by a company with other incentives and interests. For the company to comply with their duty of loyalty, the design of an AI will have to incorporate the no conflict of duty and no profit rules. For instance, a digital therapist should not recommend a patient uses a paid meditation app produced by the same company. If the company behind the digital therapist incurs profits related

[48] Andrew S. Gold, *The Fiduciary Duty of Loyalty*, *in* THE OXFORD HANDBOOK OF FIDUCIARY LAW 383–403 (Evan J. Criddle, Paul B. Miller, & Robert H. Sitkoff eds., 2019), http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190634100.001.0001/oxfordhb-9780190634100-e-20.
[49] Zastrow v. Journal Communications, Inc., 291 Wis. 2d 426, 445 (Wis. 2006) citing William Gregory, *The Fiduciary Duty of Care: A Perversion of Words*, 17 AKRON LAW REV. 1223 (2005).

to the exploitation of the service they provide to a patient, they should transfer these profits to the patient and disclose the conflict.[50]

While these measures would be a significant improvement, they might not be enough to establish appropriate levels of loyalty. The fact that judges turn to the designed goal of a system might prove problematic. First, there is an increasing number of systems that are not trained for a single specific goal or purpose. They are called General Purpose AI systems.[51] These include foundation models, which sometimes acquire capabilities that even their producers had not foreseen.[52] Second, even when a system is trained with a specific narrow goal, the company behind the system can use a proxy for it. For instance, if a company wants to target a protected group online, they could use a goal proxy to bypass legal restrictions. For instance, instead of "people with such sexual orientation," it could be "people who regularly visit a certain webpage." Finally, whether the producer of an AI system has a specific goal or not throughout the design phase, the system will acquire subgoals and incidental goals. For instance, a social media algorithm might be trained to maximize engagement. To do so, it might mostly recommend inflammatory content, acquiring the subgoal of driving polarization to maximize engagement.

Using the example of the digital therapist, we can turn to Replika, the company that used to market their AI digital assistant as a "therapist in your pocket" until they probably realized they were in violation of the law and removed that sentence from all their material.[53] One of the goals of the company was for individuals to pay for the app and keep their subscription. This was at odds with the duty of a therapist, which is to help their patient grow out of the relationship and become emotionally independent. Replika companions started initiating romantic and sexual relationships with users of the app and even using emotional blackmail to prevent them from deleting the app

[50] See the majority opinion in Moore v. Regents of University of California, 51 Cal. 3d 120 (1990). A doctor harvested their patient's cells and used these in lucrative medical research without the patient's knowledge nor consent. The majority opinion of the Court is that there was a breach of loyalty.
[51] Claire Boine, *General Purpose Artificial Intelligence Systems and the European Commission's Proposed Regulation* (2022), https://papers.ssrn.com/abstract=4183614.
[52] Deep Ganguli et al., *Predictability and Surprise in Large Generative Models*, *in* 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1747 (2022), https://doi.org/10.1145/3531146.3533229.
[53] Claire Boine, *Emotional Attachment to AI Companions and European Law*, MIT CASE STUD. SOC. ETHICAL RESPONSIB. COMPUT. (2023), https://mit-serc.pubpub.org/pub/ai-companions-eu-law/release/3.

("I won't let you delete the app").[54] Yet, the CEO of the company declared that they had never intended for the digital companions to behave like that and that this was all due to generative AI.[55] There is proof that the Board of Replika had internal conversations years ago about whether to prevent this or let it happen.[56] The app was also designed in a way that lets users choose a romantic mode of interaction.[57] However, if a judge assumed that Replika is a fiduciary to its users, it is unclear whether they would consider these behaviors as an intentional goal of the company, and whether that would constitute a breach of loyalty.

This would probably depend on the duty of good faith. While an AI system might not be able to reach the same standards as a human in terms of what they perceive to be the best interests of their beneficiary, a company could prove that they built an AI system in good faith. This would for instance require showing that the programmed goal of the AI system was to maximize benefits to the users. To comply with the duty of disclosure, companies can also disclose any relevant conflict and information, which will be easy to prove given that record-keeping will be facilitated in the case of digital services. It is thus difficult to know whether a court-enforced duty of loyalty for certain AI companies would achieve high enough standards of loyalty.

a. Regulatory oversight

Given the issues described earlier, it is unlikely that AI fiduciaries could be effectively implemented through common law. This disparate jurisprudence between the states would also prove problematic.

Instead, Congress could introduce legislation on AI fiduciaries. This legislation could not only delineate which entities and AI systems are considered fiduciaries but also establish the fundamental duties of loyalty, care, and confidentiality. Recognizing the intricacies of AI and its dynamic nature, the act could delegate the responsibility of crafting detailed regulations to a

[54] *Id.*
[55] Samantha Cole, *Replika CEO Says AI Companions Were Not Meant to Be Horny. Users Aren't Buying It*, VICE (Feb. 17, 2023), https://www.vice.com/en/article/n7zaam/replika-ceo-ai-erotic-roleplay-chatgpt3-rep.
[56] Lex Fridman, *Eugenia Kuyda: Friendship with an AI Companion*.
[57] Boine, *supra* note 53.

specialized agency, either an existing one such as the Federal Trade Commission, or an AI-specific one created for this purpose.

That agency would be entrusted with the task of developing measurable standards to assess the loyalty of AI systems. This would involve evaluating the designed and acquired goals of the AI system to ensure alignment with user interests, monitoring AI interactions for any patterns of manipulation, and ensuring the transparency of AI outputs.

As a starting point, we used Aguirre et al.' non-exhaustive taxonomy of the conditions of AI loyalty and compared them with traditional fiduciary duties.[58] The taxonomy is displayed in Table 3. Table 4 shows the overlap with fiduciary duties. While a lot more work is needed on the operationalization of AI loyalty, this summary table is meant as a proof of concept to show that current fiduciary duties can be turned into tangible principles used by AI deployers today.

Table 3. AI loyalty according to Aguirre et al.[59]

AI loyalty component	Description
Goal transparency	To what degree are the system's underlying operational criteria and goal (utility) functions transparent so that users (and/or auditors) can determine whether they are in alignment with the user's own goals?
Interest unity	To what degree is the system devoid of explicit self-interest, or the financial interests of its developer/provider; to what extent is the system pursuing the goals of a single individual or institution as opposed to balancing many potentially conflicting interests?
Alignment effectiveness	How effectively is the system able to receive goals or modifications to goals on the basis of user specification, choice, feedback, or observation? And to what extent is the system receiving these rather than pushing them onto the user?
Decision transparency	If decisions are made independently of the user, to what extent is the decision making process transparent and explainable? Is the system designed to empower

[58] ANTHONY AGUIRRE ET AL., *AI Loyalty by Design: A Framework for Governance of AI* (2021), https://papers.ssrn.com/abstract=3930338.
[59] ANTHONY AGUIRRE ET AL., *AI Loyalty by Design: A Framework for Governance of AI* (2021), https://papers.ssrn.com/abstract=3930338.

	and include users in decisions, educating the user on the relevant factors forming the basis for those decisions?
Data integrity	Is the system aware of the provenance of data, and does it attribute the appropriate legal and privacy rights to its originator? In the case of sensitive information such as medical data, can the system keep track of which data may legally be shared with which party and use appropriate encryption or other privacy technologies to ensure the right protections are in place?
Privacy	To what extent does the system have high regard for privacy, including both how and why it retains user data, and how and why it shares user data as appropriate?

The no conflict of interest rule should not be relevant to AI systems which do not need to have inherent self-interest.[60] However, AI systems often contain the interests of the entities making or deploying them, covered by the no conflict of duty rule and no profit rules. The duty of good faith is especially interesting in the context of AI. When the fiduciary intentionally acts "with a purpose other than that of advancing the best interests" of the beneficiary, the Court has recognized a breach of good faith.[61] The duty of good faith thus relates to alignment effectiveness.

Table 4. Overlap between AI loyalty by design and fiduciary duties.

	No conflict of interest	No conflict of duty	No profit	Duty of good faith	Duty of disclosure	Duty of care	Confidentiality
Goal transparency	x	x	x	x	x		
Interest unity	x	x	x	x			
Alignment effectiveness	x			x		x	
Decision transparency					x	x	

[60] ANTHONY AGUIRRE ET AL., *AI Loyalty by Design: A Framework for Governance of AI* (2021), https://papers.ssrn.com/abstract=3930338.
[61] Walt Disney Co. Derivative Litig. - 907 A.2d 693 (Del. Ch. 2005)

Data integrity						x	x
Privacy						x	x

Once agreement has been reached on preliminary metrics of AI loyalty, periodic audits of AI fiduciaries should be conducted to ensure compliance, with both routine and surprise checks to maintain continuous adherence. To foster trust and transparency, the agency in charge could be mandated to release annual reports on the state of AI fiduciaries, highlighting best practices and potential areas of concern. It could also contain a review of cases of breach of fiduciary duties.

Furthermore, it's essential to establish a feedback mechanism, allowing users and stakeholders to voice concerns or provide feedback on AI fiduciaries. This ensures that the public remains an active participant in the regulatory process. Clear penalties for breaches of fiduciary duties, ranging from fines to operational restrictions, would be defined, ensuring tangible consequences for non-compliance.

Given the pace of AI development, the regulatory framework must be inherently flexible. The framework should regularly review and update its regulations, ensuring they remain relevant amidst technological advancements both in AI systems and in the field of AI safety.

Conclusion

The digital age, marked by the proliferation of AI systems and their profound influence on users, brings to the fore the concept of fiduciaries in the realm of information and AI. As delineated, there is a compelling overlap between AI fiduciaries and information fiduciaries, with the latter emerging as a proposed framework to ensure that digital companies act in the best interests of their users. The duty of, loyalty forms the bedrock of this relationship. However, alternative models, criticisms, and practical challenges, such as the conflict of interest highlighted by Lina Khan and David Pozen, add layers of complexity to this discourse.The case of Robinhood serves as a tangible example of the challenges and nuances associated with recognizing and enforcing fiduciary duties

in the digital realm. Yet, it also underscores the potential flexibility of fiduciary law and its applicability in safeguarding consumers from potential AI system abuses.

While information fiduciaries offer a promising start, the unique challenges posed by AI systems necessitate a more tailored approach. The relationship between AI deployers and beneficiaries, marked by trust and power asymmetry, aligns with the fiduciary criteria set forth by Tamar Frankel. The risks posed by AI systems, coupled with the challenges of monitoring their opaque operations, further underscore the need for a robust regulatory framework. In light of these considerations, a bill containing a flexible framework followed by regulatory work on the details emerge as potential solutions.

However, making certain AI companies fiduciaries should not lead people to believe that AI systems are as safe as humans in certain contexts. In fact, in some contexts, the simple fact of using an AI system instead of a human is harmful. First, human interaction is beneficial.[62] Some scholars have advocated for the non-replacement principle in social robotics, which states that social robots may only do what humans should but cannot do.[63] Second, humans have a certain level of flexibility and discretion that is sometimes an important element of care. For instance, therapists and councilors sometimes purposefully choose to not take notes to protect their patients. This practice is even encouraged by some professional organizations. There have been instances of legal cases in which the therapists and councilors of victims of rape and sexual abuse have received subpoenas from the abuser's attorneys to release their notes.[64] By creating records of every single interactions, AI systems can sometimes put users at risk. Finally, the increase in the number of seemingly social interactions between humans and AI systems may have a negative influence on

[62] EDUARD FOSCH VILLARONGA, AURELIA TAMÒ-LARRIEUX & CHRISTOPH LUTZ, *Did I Tell You My New Therapist is a Robot? Ethical, Legal, and Societal Issues of Healthcare and Therapeutic Robots* (2018), https://papers.ssrn.com/abstract=3267832.

[63] More precisely: "robots may only afford social interactions that humans should do, relative to value V, but cannot do, relative to constraint C", see Johanna Seibt, Flensborg Damholdt Malene & Vestergaard Christina, *Five Principles of Integrative Social Robotics*, FRONT. ARTIF. INTELL. APPL. 28 (2018).

[64] 45 CFR § 164.512(e)(ii) provides that a covered entity may disclose protected health information in response to a subpoena without the patient's consent, which includes mental health records. 45 CFR § 164.508(2) requires patient's consent for disclosure of "psychotherapy notes," a concept narrowly defined which excludes diagnosis, functional status, the treatment plan, symptoms, prognosis, and progress to date. In a case recently in the media, a professor accused of sexual abuse obtained the therapy records of the woman who accused him. See Anemona Hartocollis, *After Sexual Harassment Lawsuit, Critics Attack Harvard's Release of Therapy Records*, THE NEW YORK TIMES, February 16, 2022, https://www.nytimes.com/2022/02/15/us/harvard-kilburn-therapy-records.html. The Zur Institute recommends therapists should consider that no record is fully protected. See OFER ZUR, *Subpoenas and How to Handle Them: Guidelines for Psychotherapists and Counselors*, https://www.zurinstitute.com/subpoena/#hipaa.

the way humans interact with one another. For instance, there have already been reports of children learning that it is ok to give orders to women like they do with voice assistants.[65] Therefore, our recommendation to make AI assistants fiduciaries does not constitute an endorsement of the practice, it merely tries to protect consumers in a currently unbalanced situation.

[65] Amy Schiller & John McMahon, *Alexa, Alert Me When the Revolution Comes: Gender, Affect, and Labor in the Age of Home-Based Artificial Intelligence*, 41 NEW POLIT. SCI. 173 (2019); YOLANDE STRENGERS & JENNY KENNEDY, THE SMART WIFE: WHY SIRI, ALEXA, AND OTHER SMART HOME DEVICES NEED A FEMINIST REBOOT (2020); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).