AI Loyalty: A New Paradigm for Aligning Stakeholder Interests

Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B. Reiner, *Associate Member, IEEE*

***Abstract*—When we consult a doctor, lawyer, or financial advisor, we assume that they are acting in our best interests. But what should we assume when we interact with an artificial intelligence (AI) system? AI-driven personal assistants, such as Alexa and Siri, already serve as interfaces between consumers and information on the Web, and users routinely rely upon these and similar systems to take automated actions or provide information. Superficially, they may appear to be acting according to user interests, but many are designed with embedded conflicts of interests. To address this problem, we introduce the concept of AI loyalty. AI systems are loyal to the degree that they minimize and make transparent, conflicts of interest, and act in ways that prioritize the interests of users. Loyal AI products hold obvious appeal for the end-user and could promote the alignment of the long-term interests of AI developers and customers. To this end, we suggest criteria for assessing whether an AI system is acting in a manner that is loyal to the user, and argue that AI loyalty should be deliberately considered during the technological design process alongside other important values in AI ethics, such as fairness, accountability privacy, and equity.**

***Index Terms*—Artificial intelligence (AI) assistant, AI ethics, conflict of interest, fiduciary duty, loyalty, value alignment.**

I. Introduction

CONFLICTS of interest can arise when we try to satisfy our duties to two or more entities. Traditionally, the parties that find themselves in such situations have been individuals or organizations. However, we are witnessing a radical shift in terms of the players that might be involved: in the modern world, artificial intelligence (AI)[1] systems introduce a new set of stakeholder dynamics where conflicts of interests arise. For instance, a user who searches for a product using an AI system might assume that the results that are returned are the most relevant, highest quality, or best value, but in fact, such systems often prioritize results that provide the most financial benefit to the software designers, and algorithmic news feeds may prioritize user engagement time and advertising dollars over users' desire for true and useful information [1]. As the capability and sophistication of AI engines providing recommendations, making decisions, and taking action mature, these misalignments in interest can be expected to grow in importance. The issue is especially problematic when we consider the power and knowledge imbalance between individual users and technology companies with financial resources greater than some sovereign governments [2]. This article aims to address the issue of conflict of interest in AI software through the lens of the concept of *AI loyalty.* While the framework that we develop may have widespread applications for AI systems more generally, we begin with the implications of AI loyalty in products that offer AI-based services to individual consumers. For this reason, we explore the concept through the particularly salient context of AI personal assistants, in which an AI system acts as an agent that makes recommendations and takes actions on behalf of a user in a way that echoes a human assistant or consultant.

II. Contemporary AI Personal Assistants

Virtual assistants powered by AI are becoming ubiquitous in the modern world. Systems, such as Apple's Siri, Amazon's Alexa, Google's Assistant, and Microsoft's Cortana, can be found on billions of smartphones, smart speakers, and digital buttons and are capable of handling a growing array of digital tasks. From the users' perspective, these assistants generally present themselves as performing tasks or providing information *for the user's convenience and benefit*, without making their limitations or conflicts of interest explicit. However, from a business perspective, these assistants may play various other roles: as a means of harvesting data that can be used to form digital profiles to inform future product development, target advertising, or feed machine learning systems, serving as a value-added to a hardware system, acting as an interface to increase product purchases, and so on [3].

Over the last 20 years, business practices enabled by technological advances in consumer tracking have changed long standing assumptions about conflicts of interest between businesses and customers. Traditionally, users purchasing a product or service were safe in assuming that the service would align with the user's interests in exchange for the financial compensation they paid. But this economic arrangement and its attendant assumptions have been undermined by the new

Manuscript received April 21, 2020; revised July 29, 2020; accepted July 29, 2020. Date of publication August 4, 2020; date of current version September 11, 2020. The work of Peter B. Reiner was supported by the Social Science and Humanities Research Council of Canada. This article was recommended for publication by Associate Editor K. Michael. *(Corresponding author: Peter B. Reiner.)*

Anthony Aguirre is with the Future of Life Institute, University of California at Santa Cruz, Santa Cruz, CA 95060 USA (e-mail: anthony@futureoflife.org).

Gaia Dempsey is with 7th Future, Santa Barbara, CA 93105 USA (e-mail: gaia@7thfuture.com).

Harry Surden is with the Law School, University of Colorado at Boulder, Boulder, CO 80309 USA (e-mail: hsurden@colorado.edu).

Peter B. Reiner is with the Department of Psychiatry, University of British Columbia, Vancouver, BC V6T 2A1, Canada (e-mail: peter.reiner@ubc.ca).

Digital Object Identifier 10.1109/TTS.2020.3013490

[1] We use "AI" here as a catchall term to include a variety of machine-learning and related techniques incorporated into contemporary systems.

class of widely available digital products, whose replication and distribution costs are negligible in comparison to that of physical goods. Today, businesses provide technological services to millions or billions of users at no direct charge and monetize this relationship through less obvious pathways [4]. These alternative pathways include the gathering and analysis of personal consumer data, aggregation and sale of that data. Firms may use that data to influence consumer purchasing, capture engagement time in order to sell ads, or other personal actions. Importantly, these monetization strategies are often not apparent to end-users. Worse, companies often take active steps to obfuscate these activities from consumers; and many of these same companies are developing the most widespread AI assistants.

This represents a potential divergence of interest that has garnered increasing attention and concern. A virtual assistant may, without disclosure, encourage (or only allow) purchases from a particular vendor or otherwise serve its home company's interests [5]. Moreover, the data-gathering business model has been heavily critiqued as being ethically fraught. Companies' financial incentive is to extract ever increasing amounts of personal information regardless of benefit to the user.[2] As a result, consumer backlash and accusations of manipulation, privacy violation, and more have been levelled at technology companies [8]–[10]. In an attempt to manage the situation, governments have imposed historically heavy fines for breaches of privacy regulations both in the U.S. [11] and in the EU where GDPR rules hold sway [12].

We propose to turn this situation on its head, beginning with an ethical view of how our relationship to AI assistants should look, from the perspective of the consumer. We will argue that a key feature of a new, ethically sound product category is that an AI assistant should be loyal to the consumer—that is, it should consider the user's interests first and foremost. Importantly, this loyalty principle applies to consumer AI systems more broadly, and not just to AI assistants. AI assistants are a specific, tractable example of a system where the AI loyalty framework can be used to shed light on analogous problems that are endemic in AI systems more generally. These issues will only become more important as these systems' capabilities increase. In Section X, we will discuss the range of factors that would need to be present in order to command a sea change in the stakeholder mindsets and ultimately widespread development and adoption of loyal AI assistants. These include market demand, regulatory pressure, consumer advocacy, and media attention.

III. What Might We Want AI Assistants to Do?

The current crop of AI assistants, even if quite limited as compared to future versions, are reasonably capable. Voice recognition software allows users to communicate with them using natural language. They can also answer factual questions, take dictation, send texts and emails, schedule events, make recommendations, and control the expanding suite of "smart" devices taking up residence in the home, car, and office.

With time, the capabilities of AI assistants are likely to increase such that they become more and more like *bona fide* executive assistants [13]. Eventually, some tasks will be carried out autonomously without user oversight and others—in particular, higher stakes scenarios that involve greater complexity—will require significant user input. Even without invoking the development of full-fledged artificial general intelligence, it seems plausible that one day AI assistants could be involved in making significant purchases [14], booking travel reservations, negotiating the purchase of a home, supporting detailed logistics, and event planning. AI assistants might also support decision-making on topics related to health, nutrition, childcare, energy consumption, job market, investments, or any other issue that an individual would ask a competent operator or highly intelligent advisor for support [15].

AI assistants could protect users from an array of cyberattacks, warding off not just spam but also phishing, malware, and even legal but exploitative scams. They could also serve as a central repository and negotiator of user preferences on privacy and similar matters, communicating those preferences to a system with which the user is interacting, and/or warning a user when a system does not respect those preferences. Similarly, users' ethical preferences (for example, to favor or avoid particular industry practices or company qualities) could be folded into recommendations to the user.

On a more aspirational level, an AI assistant could help users achieve a better state of well-being—a software layer that users can cooperate with to help them more effectively steer, drive, and course-correct along their personal path of self-improvement. Developers are already releasing AI assistants tasked with helping users improve their health and material welfare, gently nudging [16] them toward better nutrition, serving as AI therapists, and helping them save for a rainy day [17]. Even more ambitious would be programming that helps users further develop the skills that provide meaning to the human experience: the satisfaction that derives from accomplishment, the value of social connectedness, the nourishment of one's inner life. AI assistants could "know" when and how it is more appropriate to encourage the user to reflect on the values underlying their habits and teach them to evaluate their lives' competing demands [18]. Designing AI assistants with these features would go some distance in lessening the worry that over-reliance upon AI assistants for instrumental tasks might enfeeble us, diminishing our capability to effectively navigate the world [19]. The concept of AI loyalty becomes of paramount importance in systems with such a high degree of access to the user's sense of identity, well-being, and general mental and emotional state.

IV. Aligning Interests—The Case for Loyal AI Assistants

For most people, using an AI assistant, such as Siri or Alexa, simply involves speaking into a device and hearing a response. While the small print in the terms of use generally

[2] Such systems can act not only to extract data but also to deliberately increase reliance on the ecosystems of products they connect with. Because of this, negative feedback loops can result that decrease the agency of the individual yet lock them into an ecosystem that is all but necessary for their participation in large swathes of the global economy and social fabric; see [6], [7].

make it clear that collecting and exploiting user information is a condition of accessing the service, most users ignore the terms of use entirely [20]. The relatively unfettered collection of user information, sets the stage for AI-driven algorithms to persuade, cajole, and manipulate user behavior [9], [21]—all without much incentive to consider the best interests of the user. We suggest that loyal AI assistants—AI agents that explicitly put the interests of the user at the fore—represent an appropriate resolution of this problem. Such AI loyalty may also confer a significant marketing advantage, provided they are implemented in a trustworthy manner.[3] Loyalty of this sort represents a version of the value alignment problem—the challenge of ensuring that an AI-driven agent act in accord with some set of values. We will not recapitulate the full set of arguments that have been proffered on this topic but note that its solution is fundamental to the development of the next generation of AI systems [22], [23].

The idea of "loyalty" is well established in human professional relationships in the form of the fiduciary duty owed to their clients by physicians, therapists, and lawyers [24]. We expect these fiduciaries to put our interests first and to carefully and ethically handle any conflicts of interest that may arise. In turn, we feel comfortable sharing important and personal information and trusting their recommendations. AI loyalty would logically be even stronger than that of a human fiduciary, as unlike a doctor or lawyer, the AI presumably has no interest (for example, salary, pride, etc.) of its own.[4] The well-being of the end-user could be absolutely paramount in terms of recommended products or purchases. The high ethical standards that the law places upon fiduciaries might be appropriate to apply with even greater expectations to the loyalty of an AI system.

In terms of actions, depending upon the business model employed (see below), a loyal AI assistant could be configured such that it was primarily responsive to the interests of the user so long as actions that it might be asked to provide are not illegal. (Even here exceptions could obtain. Imagine a scenario in which one is rushing to the hospital in a self-driving car, and the user asks the AI assistant to go 5 miles per hour above the speed limit. The AI assistant might respond by reminding the user that speeding is against the law, and the user could accept responsibility for the legal infraction. If such exceptions become common, rules about AI behavior may become more pastiche than consensus.)

A well-functioning loyal AI assistant—preferably functioning in the context of a clear and robust legal and regulatory framework—would reinforce the possibility of trust between user and algorithmic system. Indeed, it would seem appropriate for AI loyalty to join the existing pantheon of attributes of trustworthy AI, such as reliability, explainability, privacy, and fairness [25]. That an AI system is capable of loyalty might be part of a chain of trust [26], being included in a fact sheet accompanying the technology [27], as well as ongoing monitoring in what has been termed a human impact assessment for technology [28]. Assuming that the duty of loyalty is maintained over time, one can imagine the relationship moving toward the thick sort of trust that develops healthy interpersonal relationships, in which the parties have a shared history that reinforces mutual confidence in their actions [29].

As with interpersonal relationships, inappropriately applied trust could be problematic with respect to AI assistants [30]. The danger is that we wrongly trust and uncritically accept information from an algorithmic agent without appropriate safeguards being in place. This is precisely why we need a means of credentialing AI loyalty, so that we can have confidence in the trust that we place in AI assistants.

This is important not just on a social level but also to improve the overall function of interactions between systems. The deeper the trust that users ascribe to their loyal AI assistants, the more freely they will share information with it. As a result, the loyal AI assistant will have more robust information at its (virtual) fingertips and therefore be better able to discharge its duties toward the user. This virtuous circle is exactly the opposite of the vicious circle we see underway in the current regime, where people who distrust social media are encouraged to be careful about what they share [31], resulting in less robust images of who they are. Trust allows the AI system not just to receive information but also to carefully, appropriately, and where applicable, anonymously share information, opening the possibility of new functionalities. Modern anti-spam systems, for example, overpower spam generators by sharing information about what is labeled as spam [32]; loyal AI assistants, similarly, could help their user coordinate with others with common interests. This could include identifying and mitigating low-value drains on users' time and money, as well as protecting the user against manipulative tactics. A loyal AI could also be entrusted to identify (and if so requested, take advantage of) opportunities for its user that are genuinely in the user's interest.

More broadly, our modern Internet-mediated economy embodies an unprecedented asymmetry of information, in which, in many cases, individuals cannot possibly make rational informed decisions reflective of their own interests, values, and objectives without a level of time, effort, and expertise that is unavailable to the vast majority of people. Loyal AI assistants could help redress this imbalance by providing individuals with trusted tools that can analyze and navigate a high-complexity economic environment, as well as allow coordinated action to counteract the power of corporations and nation states.

A well-established system of loyal AI assistants would also have significant wider social implications. It is already known that loyalty promotes trust in data stewards [33]. In order to garner the trust of users, loyal AI assistants would need to satisfy other features, including at a minimum, the ability to

[3] It might be objected that conflicts of interest between AI users and developers should simply be transparently declared, and that this information is sufficient for users to fold into their decision making. But in a system that has such conflicts built-in with some requirement to declare them, the incentives most naturally lead to the declarations either being hidden in lengthy, abstruse, universally unread terms-of-use agreements, or—where this is disallowed—so ubiquitous as to be simply wasteful and dismissed without effect, as in endless GDPR notices. Neither case leads to meaningful mitigation of the conflict.

[4] We leave aside the issue of whether future advanced AI systems could meaningfully have their own interest but suggest that even if it were possible to create such systems, AI assistants should probably not be among them.

explain their actions in easily understandable terms, to maintain data security, and, importantly, be able to communicate with outside entities and agents without compromising privacy. Models for the latter include the well-developed examples of differential privacy [34] and contextual integrity [35], or the federated approach of the privacy-preserving open-source AI assistant framework Almond [36]. The idea of AI loyalty also aligns closely with that of "human compatible" AI [22], in which AI systems aim to accomplish human goals rather than their own (even if human-provided) goals.

While loyalty in many ways may appear an unalloyed good, there are many subtleties, as in all human interactions. For example, how should the user's interests be balanced against interests other than those of the AI (which presumably do not exist) or its parent company? An objection to the loyalty requirement has been raised by AI pioneer and practitioner Stuart Russell who rightly points out [22], in the context of quite advanced AI assistants, that an AI might follow the law but still unscrupulously trespass social norms in the process of executing its loyalty obligations to its user. He posits a situation in which the user has accidentally double booked for dinner—both with his wife for their 20th anniversary and the secretary-general who is flying in for the occasion. As a solution to the dilemma, the AI causes the secretary-general's plane to be delayed. Russell's solution is for the AI to follow a set of moral rules known as consequentialism. We suggest that true loyalty means reflecting the user's values, their social etiquette, and their general preferences about navigating the world. Presumably, the user, in this example, would not have wanted this drastic consequence, and if the AI system had truly been following the user's values, it would have arrived at a more palatable solution. For simple AI assistants, such as those available today, such regard would need to be instantiated directly into the system, implicitly consented to by the user when adopting the system. More advanced systems with a greater action space could include, for example, user-definable settings that characterize how to weigh the user's interest against others', with some required minimum weight accorded to the latter.

In the future, more sophisticated systems might have a detailed model of their users' preferences and could learn from their user's decisions how to accord weight to the interests and preferences of others, with some hard ethical limits to the loyal AI assistant's behavior in place. That is, an advanced loyal AI assistant should include the consideration of others in its recommendations and actions specifically because its user would. However, this arrangement would do nothing to mitigate the amoral actions of bad actors. Such individuals would likely use the power of AI to further their amoral interests—this is an inevitable consequence of putting the growing power of AI at the service of people who enjoy significant liberty in their actions. New legal and social norms will develop to address this reality, often driven by notable/newsworthy transgressions. On the other hand, in the case of people *trying* to do the right thing, AI assistants with such advanced capabilities could even help better align the actions of the user with the user's own moral compass: anytime the user makes a request that appears to transgress what the loyal AI assistant understands to be the morally appropriate thing to do, it could prompt the user to reflect on the issue and await further instructions. Small nudges, such as this, are already in place in the AI that monitors offensive posts on Instagram [18] and could easily be expanded upon and deployed more widely.

V. Relation to the Issue of AI Alignment

The alignment of AI assistants with their users' interest is a special case of a wider problem. As AI systems become more capable it will make sense to cede more decisions to them. But how can these decisions then be assured to be consonant with the goals and values of their operators, so that it makes sense for an individual (or organization) to delegate these decisions or trust these recommendations? Numerous AI experts among others have pointed out the profound and fundamental difficulty of this problem [22], [37], [38] arguing that any explicitly specified set of objectives for an agent operating over a very wide action space are virtually guaranteed to lead to unforeseen negative side effects. Just as with human assistants, AI assistants will make mistakes—both in failing to do what they are trying to do, and in trying to do the wrong thing. This will make AI assistants challenging as products, because many users employing different and customized AI assistants for various tasks will lead to instances of users pushing AI assistants in potentially problematic directions, and instances of users objecting to (i.e., complaining about) things the assistants do. This, however, is an opportunity as well, since through this process the framework of the assistants and their design underpinnings will be repeatedly and adversarially tested and improved, starting when the stakes are low. This seems much more likely led to robustly aligned AI architectures than for AI systems with limited interaction with a limited number of users and may very well significantly contribute to the broader challenge of AI alignment.

The type of loyalty discussed here is not the only form of AI alignment. Loyalty as such need not be to an individual person. We may imagine AI systems designed to be loyal to a group, a legal entity, such as a corporation or government, a nation, or even to humanity as a whole. As with human affairs, it may be challenging to determine how to aggregate the preferences of multiple entities within the group, and in those cases the label of loyalty would fit less well, even for systems that are explicitly designed to behave according to agreed upon moral tenets.

A closely related issue is that of *responsibility:* if an action is taken by an AI system designed and built by a provider, and used by a person, who is responsible for that action? In the AI loyalty paradigm, the AI system is operating at the behest and behalf of its user in accord with its design and creation by the provider; this largely removes the possibility of placing that responsibility in the AI system. This seems appropriate, as—at least in the foreseeable future—AI systems cannot really be *held* responsible in a legal or moral sense. Avoiding decoupling of decision and responsibility is thus a feature and a core part of the AI loyalty paradigm, which aims for decisions to be *delegated* to AI systems, but not *ceded* to them.

VI. Criteria for Near-Future Loyal AIs

Although some loyal AI capabilities will require future technical advances, this framework applies to contemporary AI assistants. Here, we sketch some principles and criteria for designating an AI assistant system "loyal."

A. Value Alignment

To the extent that the system acts independently of the user to facilitate the completion of some action or task.

1) The system should be deliberately designed to eliminate clear conflicts of interest—e.g., AI systems taking automated actions or providing prioritized recommendations that financially benefit the creators or funders of the AI system.
2) When clear conflicts of interests do exist, the system should transparently and saliently indicate to users the presence of such conflicts.
3) The system's underlying operational criteria and goal (utility) functions should, at minimum, be made transparent so that users (and/or auditors) can determine whether they are in alignment with the user's own goals.
4) Preferably, the criteria and goals should be adjustable in terms of balancing tradeoffs in the prioritization of optimizable factors (e.g., price, speed, cost, privacy, etc.).
5) Optimally, the values utilized by loyal AI should be derived from revealed preferences [23], learned directly from the user where possible,[5] appropriately and efficiently requesting user input as needed.

Maintaining value alignment between an AI system and its user is a nontrivial problem [40], [41] but also an area of active investigation [23], [42]–[44]. As discussed below, AI assistants may be a useful proving-ground for alignment techniques.

B. Decision Transparency

To the extent that decisions are made independently of the user, the decision-making process should be transparent and explainable. The system should be designed to empower and include users in decisions, educating the user on the relevant factors forming the basis for those decisions. This accomplishes three important objectives: 1) it conditions and trains users in the powers and limitations of the system [45]; 2) it protects against a form of learned helplessness that has been termed digital resignation [46]; and 3) it facilitates human intervention in the case of automated decision failures.

C. Data Integrity

The system should be aware of the provenance of data and attribute the appropriate legal and privacy rights to its originator. For example, the system should be capable of tracking the origin of data that is being monetized and directing payments (or other value commensurate with that of the data) to the correct entities. Or, in the case of sensitive information, such as medical data, the system should be capable of keeping 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.

D. Personal Privacy

The system should have extreme regard for privacy, including both how and why it retains user data, and how and why it shares user data as appropriate (to accomplish a given task, for example, in an anonymized and encrypted method allowing network effects with other AI assistants.) In addition, the system makes explicit to the user the privacy risks of any particular action, when appropriate.

VII. Legal Framework

To what extent, if any, should the law play in encouraging or requiring AI loyalty? One possible approach involves statutory or administrative regulation. Congress, state governments, or regulatory agencies could introduce legal rules requiring the creators of AI systems to incorporate AI loyalty principles in various aspect of their product lifecycle [47] For example, legislation could mandate that AI systems disclose and make transparent to users obvious and clear conflicts of interest. By way of comparison, at present, there are Federal Trade Commission rules that require bloggers and others with an Internet presence to disclose any conflicts of interests involved in products or services that they endorse [48]. Similarly, we could imagine analogous transparency regulations in the context of AI systems requiring them to disclose any clear conflicts of interest from automated actions on behalf of a user.

Relatedly, AI loyalty could be viewed through a more general consumer protection lens. Arguably, AI systems that appear to produce accurate and relevant actions for users, but that are actually fostering the interests of others (e.g., subtly promoting the AI-creator's own products at the expense of more relevant or less expensive products of competitors), could be characterized in terms of deceptive or unfair trade practices. With this legal framing, there already exist numerous agencies that could implement AI loyalty principles through current law, without necessarily creating new rules [49]. Such consumer protection agencies use existing law to advance AI loyalty through enforcement actions against the creators of AI systems that engage in deceptive and nontransparent actions or produce recommendations involving nonbeneficial conflicts of interest against consumers.

A different AI loyalty regulatory approach might focus upon clear and transparent explanations for users as to how automated results were produced even in the absence of direct conflicts of interest. Requiring AI systems to detail the steps and data sources that led to a particular recommendations or actions might act as a check on unfavorable, automated actions. We see analogs to such requirements in the privacy realm, where recent regulation from California (the California Consumer Privacy Act) and the European Union [The General Data Protection Regulation (GDPR)] contain analogous rules with respect to privacy and personal data [50]. This approach seems reasonably promising, as there has been much discussion (and proposed legislation) at the

[5] This fits the existing machine learning framework of Inverse Reinforcement Learning; see [39] and, addressing longer term issues [22].

state- and federal-level concerning the regulation of AI more broadly [47].

Another possible regulatory route is to focus on the product development cycle itself. We could imagine rules requiring the developers of AI systems to explicitly consider issues of AI loyalty as part of the larger development process. By way of comparison, the GDPR requires companies to consider privacy issues during the development stage of technological systems [51]. It is not hard to imagine an analogous requirement that "AI Loyalty" principles also be considered, alongside principles of privacy, during AI system development.

Contract law might also play a role in promoting AI loyalty. In some cases, vendors or users of AI assistants might promulgate explicit terms of service or policies governing their use and behavior. In certain circumstances, one could imagine explicit contractual promises that AI assistants will make reasonable efforts to implement systems that promote loyalty, privacy, explainability, and data security. However, unless specifically required by law, such explicit contractual promises do not appear to be a promising avenue due to the difficulty in defining these standards, the subjective nature of these criteria, and the risk of abuse by opportunistic litigants.

In the U.S., the courts may also have a regulatory role in developing common-law liability for AI systems that engage in user disloyalty or create harm to others. One could imagine a body of AI loyalty rules slowly developing in such a common law approach through individual litigation or consumer class action lawsuits. The creation of common law duties of implied AI loyalty, arising out of caselaw is certainly feasible, much in the way that consumer goods now have common law implied warranties of merchantability even in the absence of explicit warranties. Courts might, for instance, impose loyalty duties on creators of AI systems that echo some of the fiduciary requirements seen in the professional context of attorneys and doctors. Such court-created fiduciary duties might mandate that user-facing AI systems be designed to act in the best interests of their users, or at least to avoid clear conflicts of interest that harm users. [52] Other concerning scenarios involve AI systems whose automated actions cause external harms to third parties other than the users. There courts might employ existing negligence/duty of care doctrines where the creators of AI systems, or the users of AI, pose unacceptable and unreasonable risks to third parties [53].

To a greater or lesser extent, all government regulation will have to confront a series of difficult but central issues: how will laws coherently define concepts, such as "conflicts of interest," "best interest of user," or "loyalty?" In other words, will regulators be able to provide legal standards for such abstract concepts that can be applied in a reasonable, efficient, workable, and effective manner? This, in turn, raises other difficult topics. How do we actually determine or measure a particular user's "best interest?" What if an individual user has multiple interests that benefit her, but that conflict with one another? How does a system choose among multiple, competing interests, all of which may plausibly be said to benefit that user? What if an automated action benefits a user and also someone else, such as the AI-creator? What if there are multiple users effected by an automated action? Whose user's interests are we measuring? What if one user's best interests actually harm others and whose should take priority? These critical open questions around the intersection of loyalty in AI will need to be resolved from both an ethical and a legal perspective.

These are difficult issues, but we can certainly distinguish these harder issues from other scenarios that may be clearer and more egregious, and which could be regulated today. At present, the law could readily target issues of clear "AI disloyalty"—those scenarios in which AI systems act against users, and in favor of another (i.e., a system that makes a subpar automated decision the expense of a user in exchange for a financial kickback for an AI creator) – without necessarily having to immediately solve the broader, more philosophically challenging questions of "AI loyalty" more generally. In other words, it is possible to fruitfully regulate clear "AI disloyalty" scenarios in the near term even if lawmakers need more time to address more value-laden, and abstract AI loyalty topic such as contexts where user interests are simply under-specified, or are not maximally satisfied by AI systems.

Finally, altogether different approaches to AI loyalty might focus upon limited (or no) government involvement. One could imagine a form of industry self-policing that might arise in which the corporate creators of AI systems voluntarily develop and implement AI loyalty best practices during the process of technological design. The possibility of industry self-regulation is not completely far-fetched. There have been past examples of such self-regulatory efforts in the area of privacy and data anti-discrimination. These voluntary, nongovernmental regulatory efforts have arisen largely as a reaction to public criticism and as a means to head off pending government regulation [54].

Relatedly, AI principles might be fostered through "market discipline." Market discipline is the concept that economic competition (as opposed to government regulation) can bring about some desired social goal when implementing that goal actually confers a competitive advantage [55]. In this vein, one could imagine consumer rights groups desiring or demanding that consumer AI systems become transparent, free of conflict of interest, and loyal to user interests. If such consumer demand were to arise, firms could voluntarily implement and advertise AI loyalty as a market distinguisher for their products. That could result in increased market share and economic advantage over competitors who produce AI products with embedded conflicts of interests, producing competitive pressure that could cause AI loyalty to diffuse more broadly.

However, such a "market-discipline" scenario might not actually transpire, as it would require both consumers and producers to value AI loyalty (over other features) such that it would shift their usage preferences in a meaningful way. By contrast, history has shown that consumers often prefer other factors such as speed, ease of use, or low or free price, even at the expense of the concepts such as transparency or privacy that are related to AI loyalty—or consumers may not recognize the importance of these factors until their absence is so baked-in to large-scale systems as to be both problematic and difficult to change. Thus, while loyalty may be a significant product differentiator, this market discipline driver is unlikely to suffice on its own, without also involving explicit government regulation, to create the conditions necessary for widespread adoption [55].

In sum, while industry self-regulation and market discipline may play a part in fostering AI loyalty, they are unlikely to sufficiently advance the concept on their own. Rather, some sort of government regulation will likely also have to play a role to ensure that AI loyalty principles are implemented such that they sufficiently protect users and advance trust.

VIII. Paying for Loyal AI Assistants

The concept of loyal AI builds upon, but is distinct from the proposal that large technology companies ought to be regulated as information fiduciaries [56]. Indeed, the information fiduciary model has been critiqued for condoning the tech titans in maintaining their existing business model, collecting user data to drive microtargeting of advertising [57], thereby perpetuating conflicts of interest. Loyal AI demands modification of this business model. At the same time, loyalty could be a major product differentiator and competitive advantage, particularly as privacy and related concerns continue to grow and potentially come under stronger regulatory sway.

It is likely that any model in which the AI system provider has a financial model with revenue resulting from the way the loyal AI assistant operates is likely to lead to conflicts between revenue maximization and loyalty to the user. This conflict can be removed if the revenue is tied to the fact that the assistant is used, rather than what it actually does.

One option, therefore, would be for a loyal AI to be bundled with another product, consistent with a business model in which “turning the flywheel” represents a path to success [58]. For example, Apple’s AI assistant Siri is incorporated into all iPhones. Offering a loyal version of Siri would represent an added inducement for people to purchase iPhones, with Apple incorporating the cost of the AI functionality and ongoing support and development in the purchase price of the device and their ongoing support and service plans, such as Apple Care and iCloud. Whether introduced by Apple or its competitors, one can imagine a scenario in which a loyal AI assistant would confer significant competitive advantages to its producers. If successful, such a consumer product would create social and market pressure that could modify the unhealthy balance of power that exists today in the industry.

An alternative model might involve subscription pricing. Basic versions of a loyal AI assistant could be offered for a modest cost, with capability enhancements available as the equivalent of in-app purchases. These are only the most obvious solutions, but the power of entrepreneurial ideas will undoubtedly lead to alternative approaches, each competing for market share in their own way. Some of these business models may come with their own problems—for example, there is the worry that if AI assistants with a wide range of capabilities are available at a wide range of cost points, loyal AI assistants might be available most easily only to those with the means to pay for them, widening the gap between the haves and the have-nots.

IX. Power Asymmetries and Coordination

Our modern Internet-mediated economy embodies an unprecedented asymmetry of information, and in some ways power, between large companies that can collect and process information from many people and enact decisions that affect many people at once, and individuals who must each make decisions and take action based on their own information. Governments and the legal system (including class-action lawsuits) can and do protect citizens from physical harm and (at some level) fraud etc.; but what protection is available to people against pervasive manipulation, privacy invasion, false (and too-numerous) choices, etc.? The traditional model—successful in many ways—has been that informed consumers and citizens “vote with their wallet” and target their resources toward companies and products that serve them best. But the modern information economy is arguably so complex and information-asymmetric that individuals cannot possibly make rational informed decisions reflective of their own interests, values, and objectives without a level of time, effort, and expertise that is unavailable to the vast majority of people. This trend is unlikely to change unless a new ingredient is added that helps tip the balance back in favor of autonomous individuals, by providing them with trusted tools that can analyze and navigate a high-complexity economic environment, as well as allow coordinated action to counteract the power of corporations and nation states. Loyal AI assistants are an example (but only one) of such a new ingredient.

X. Recommendations

The aim of this article has been to bring attention and conversation to the importance of loyalty—or lack thereof—in current and future AI systems, focusing on the example of AI assistants. To the extent that loyalty is desirable, what could make it the norm? Commonly, issues of public concern (e.g., climate change) are only considered by powerful economic players when a combination of market incentives, government regulation, consumer demand, consumer advocacy, media pressure, and shareholder activism are present—no single ingredient alone is sufficient to catalyze meaningful, widespread change in corporate operations and policy. Thus, we offer a diverse set of initial recommendations toward the end of AI loyalty.

1) AI researchers and developers should continue to devise, develop, and prototype systems and protocols necessary for loyal AI systems, including private information storage and appropriate sharing, data provenance tracking, decision and goal transparency, and conflict of interest monitoring.
2) AI companies in or adjoining the market of personal assistants should consider loyalty as a potential design feature for their existing products and consider developing products with loyalty at their core.
3) Major tech companies offering personal assistants should consider loyalty alongside properties like privacy and discrimination. Moreover, they should create internal processes that empower technologists and other product development specialists to explicitly query the

issue of conflict of interest throughout the development pipeline. Companies for whom AI loyalty is a natural part of their product offering should leverage this fact for market advantage.

4) Policymakers should consider loyalty alongside other pro-social AI system attributes, such as transparency, nondiscrimination, privacy, and safety. At a minimum, transparency and disclosure of conflicts of interest should be strongly considered as part of any set of strictures placed on consumer-facing AI products.
5) Consumers, while demanding appropriate levels of effectiveness, privacy, safety, and fairness from the products they use, should consider how important AI loyalty is to them and exhibit this both explicitly and in purchase/use preferences.

XI. Conclusion

The past 20 years have seen the explosive growth in the power and capability of online platforms that mediate the connection between users and nearly every important part of the social, economic, intellectual, and even natural world. The power and reach of these platforms have created enormous convenience for many users and increasingly are becoming an indispensable part of modern life. However, their advertising-focused ethos and the nearly regulation-free model in which they have been developed "fast" without concern for "breaking things" has resulted in a system with fundamental drawbacks.

Many of these drawbacks are embodied in current-day AI assistants. There are many instances in which AI systems appear to be acting in the users' best interests, but in fact do not do so. Rather, these systems are subject to either deliberate or accidental technical designs that promote the interests of the system creators (or others) to the detriment of users. The absence of what we have called "AI loyalty" has largely flown under the radar—as compared, for example, to privacy and equality—for two reasons. First, although AI systems make automated decisions on behalf of lay users today, these automated decisions tend to be fairly inconsequential data retrieval tasks, such as playing music, creating task reminders, or returning results for search queries about product purchasing or navigation. Second, the technology underlying contemporaneous AI systems tends to be relatively limited, thereby limiting their scope and reach.

This will change. When an AI system with an embedded conflict of interest subtly promotes one less attractive product at the expense of another, the impact is relatively small. But if one considers the widespread adoption of AI in the analysis of job interviews, in the automated assessment of banking and credit applications, in medical diagnostics, and as a (widely derided) means of parole assessment, it becomes clear that AI systems are rapidly being implemented in high-stakes arenas. It is only a matter of time before AI systems develop sufficient technological capabilities that they will be assisting consumers with more consequential decisions, such as assisting parents in finding the best daycare centers for their children, or helping doctors prescribe medication, or managing wealth portfolios independent of human input for long periods of time. These are precisely the scenarios for which we must avoid conflicts of interest. It is, therefore, important to open up the conversation into AI loyalty today, while the stakes and capabilities remain comparatively low, and to explicitly incorporate considerations of AI loyalty into the technological design and deployment process. In this article, we have outlined criteria to be considered in doing so.

We have also emphasized that beyond mitigating conflicts of interest that are detrimental to users, AI loyalty presents an opportunity: genuinely trustable AI systems can provide services—for example, using highly private or proprietary data—that could not (or at least certainly should not) be provided by potentially disloyal ones.

In the years to come, the role of AI-based software agents in the affairs of humans will grow very substantially. We suggest that the current trajectory in which AI systems grow increasingly capable and versatile while harvesting (and selling) immensely detailed user profiles, providing opaquely sourced and reasoned recommendations, and acting in the interest of the user only insofar as it aligns with the overall growth and profit goals of the system's parent company runs the risk of leading to ever more dystopian outcomes. We can also imagine, however, a trajectory in which individual autonomy is dramatically empowered and overall well-being significantly improved, by loyal AI systems that users can, will, and should trust.

Acknowledgment

The authors thank Jared Brown, Richard Mallah, Adrian Byram, and Imre Bárd for comments on an earlier version of this article.

References

[1] Z. Tufekci. (Apr. 22, 2019). *How Recommendation Algorithms Run the World. Wired.* Accessed: Jan. 19, 2020, [Online]. Available: https://www.wired.com/story/how-recommendation-algorithms-run-the-world/

[2] P. Khana. (Mar. 15, 2016). *These 25 Companies Are More Powerful Than Many Countries: Going Stateless to Maximize Profits, Multinational Companies Are Vying With Governments for Global Power. Who Is Winning? Foreign Policy*. Accessed: Mar. 3, 2020. [Online]. Available: https://foreignpolicy.com/2016/03/15/these-25-companies-are-more-powerful-than-many-countries-multinational-corporate-wealth-power/

[3] C. Horgan. *We Already Know What Our Data Is Worth. OneZero*. Accessed: Jun. 26, 2019. [Online]. Available: https://onezero.medium.com/we-already-know-what-our-data-is-worth-48bca5643844

[4] J. Lariner, *Who Owns the Future?* New York, NY, USA: Simon and Schuster, 2014.

[5] J. Del Rey. (Dec. 3, 2019). *The Best Cyber Monday Deals According to Alexa: Any Amazon Owned Brand. Vox*. Accessed: Jan. 20, 2020. [Online]. Available: https://www.vox.com/recode/2019/12/3/20992885/best-amazon-cyber-monday-deals-alexa-private-label-brands

[6] G. Maldoff and O. Tene, "The costs of not using data: Balancing privacy and the perils of inaction," *J. Law Econ. Policy*, vol. 15, pp. 41–66, 2019.

[7] A. Siemoneit, "An offer you can't refuse—Enhancing personal productivity through 'efficiency consumption," *Technol. Soc.*, vol. 59, Nov. 2019, Art. no. 101181. [Online]. Available: https://doi.org/10.1016/j.techsoc.2019.101181

[8] B. Frischmann and E. Selinger, *Re-Engineering Humanity*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[9] D. Susser, B. Roessler, and H. Nissenbaum, "Technology, autonomy, and manipulation," *Internet Policy Rev.*, vol. 8, pp. 1–22, Jun. 2019.

[10] S. Zuboff, *The Age of Surveillance Capitalism*. New York, NY, USA: Public Affairs, 2019.

[11] *Federal Trade Commission, FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*. Accessed: Jul. 24, 2019. [Online]. Available: https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions

[12] *Alpin, Major GDPR Fine Tracker—An Ongoing, Always-Up-To-Date List of Enforcement Actions*. Accessed: Feb. 16, 2020. [Online]. Available: https://alpin.io/blog/gdpr-fines-list/

[13] M. J. Duncan, "The case for executive assistants," *Harvard Bus. Rev.*, vol. 89, pp. 81–91, May 2011.

[14] V. Lukosius and M. R. Hyman, "Personal Internet shopping agent (PISA): A framework," in *Proc. Atlantic Market. Assoc.*, 2018, pp. 25–36.

[15] P. Pico-Valencia and J. A. Holgado-Terriza, "Agentification of the Internet of Things: A systematic literature review," *Int. J. Distrib. Sensor Netw.*, vol. 14, pp. 1–20, Oct. 2018.

[16] R. H. Thaler and C. R. Sunstein, *Nudge*, Penguin, Park Imperial, NY, USA, 2008.

[17] Z. Murphy. (May 24, 2019). *Dan Ariely on How Qapital Uses Behavioral Finance Principles to Help People Save More. Tearsheet*. Accessed: Jan. 20, 2020. [Online] Available: https://tearsheet.co/new-banks/dan-ariely-on-how-qapital-uses-behavioral-finance-principles-to-help-people-save-more/

[18] L. S. Sullivan and P. B. Reiner, "Digital wellness and persuasive technologies," *Philos. Technol.*, to be published. [Online]. Available: https://doi.org/10.1007/s13347-019-00376-5

[19] N. Carr, *The Glass Cage*. New York, NY, USA: W. W. Norton, 2014.

[20] J. A. Obar, "The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services," *Inf. Commun. Soc.*, vol. 23, pp. 128–147, Aug. 2018.

[21] S. C. Matza, M. Kosinski, G. Navec, and D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proc. Nat. Acad. Sci.*, vol. 114, pp. 12714–12719, Nov. 2017.

[22] S. Russell, *Human Compatible*, Los Angeles, CA, USA, Viking, 2019.

[23] I. Gabriel. (2020). *Artificial Intelligence, Values and Alignment*. [Online]. Available: https://arxiv.org/abs/2001.09768

[24] A. S. Gold and P. B. Miller, *Philosophical Foundations of Fiduciary Law*. Oxford, U.K.: Oxford Univ. Press, 2014.

[25] (2019). *High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[26] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel. (2019). *The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies*. [Online]. Available: https://arxiv.org/abs/1912.00782

[27] M. Arnold *et al.*, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM J. Res. Develop.*, vol. 63, pp. 1–13, Sep. 2019.

[28] R. A. Calvo, D. Peters, and S. Cave, "Advancing impact assessment for intelligent systems," *Nat. Mach. Intell.*, vol. 2, pp. 89–91, Feb. 2020.

[29] F. Niker and L. S. Sullivan, "Trusting relationships and the ethics of interpersonal action," *Int. J. Philos. Stud.*, vol. 26, no. 2, pp. 1–14, 2018.

[30] P Andras *et al.*, "Trusting intelligent machines: Deepening trust within socio-technical systems," *IEEE Technol. Soc. Mag.*, vol. 37, no. 4, pp. 76–83, Dec. 2018.

[31] F. Brunton and H. Nissenbaum, *Obfuscation: A User's Guide for Privacy and Protest*. Cambridge, MA, USA: MIT Press, 2015.

[32] P. Gillin. (Nov. 2, 2016) *The Art and Science of How Spam Filters Work. Security Intelligence*. Accessed: Jan. 21, 2020. [Online] Available: https://securityintelligence.com/the-art-and-science-of-how-spam-filters-work/

[33] N. Richards and W. Hartzog, "Taking trust seriously in privacy law," *Stanford Technol. Law Rev.*, vol. 19, pp. 431–472, Sep. 2016.

[34] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Differential privacy—A primer for the perplexed," in *Proc. Conf. Eur. Statist. Joint UNECE/Eurostat Work Session Statist. Data Confidentiality*, 2011.

[35] H. Nissenbaum, *Technology, Policy, and the Integrity of Social Life*. Stanford, CA, USA: Stanford Univ. Press, 2009.

[36] M. S. Lam, G. Campagna, S. Xu, M. Fischer, and M. Moradshahi, "Protecting privacy and open competition with Almond: An open-source virtual assistant," *Crossroads*, vol. 26, no. 1, pp. 40–44, 2019.

[37] M. Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY, USA: Knopf, 2017.

[38] N. Bostrom, *Superintelligence*. Oxford, U.K.: Oxford Univ. Press, 2014.

[39] S. Arora and P. Doshi. (2019). *A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress*. [Online]. Available: https://arxiv.org/abs/1806.06877

[40] R. Bensinger and S. Russell. (Nov. 26, 2014). *AI Value Alignment Problem Must Be an 'Intrinsic Part' of the Field's Mainstream Agenda. Less Wrong*. Accessed: Feb. 16, 2020. [Online]. Available: https://www.lesswrong.com/posts/S95qCHBXtASmYyGSs/stuart-russell-ai-value-alignment-problem-must-be-an

[41] S. Russell. *Provably Beneficial Artificial Intelligence—EECS at UC Berkeley*. Accessed: Feb. 16, 2020. [Online]. Available: https://people.eecs.berkeley.edu/∼russell/papers/russell-bbvabook17-pbai.pdf

[42] P. Eckersley. (2019). *Impossibility and Uncertainty Theorems in AI Value Alignment*. [Online]. Available: https://arxiv.org/abs/1901.00064

[43] F. Rossi and N. Mattei, "Building ethically bounded AI," in *Proc. AAAI*, 2019, pp. 9785–9789.

[44] R. Noothigattu *et al.*, "Teaching AI agents ethical values using reinforcement and policy orchestration," in *Proc. IJCAI*, 2019, pp. 6377–6381.

[45] D. C. Engelbart, "Augmenting human intellect: A conceptual framework," Inst. Directorate Inf., Stanford Res. Inst., Menlo Park, CA, USA, Rep. AFOSR-3233, 1962.

[46] N. A. Draper and J. Turow, "The corporate cultivation of digital resignation," *New Media Soc.*, vol. 21, no. 8, pp. 1824–1839, 2019, doi: 10.1177/1461444819833331.

[47] R. Calo, "Artificial intelligence policy: A primer and roadmap," *UC Davis Law Rev.*, vol. 51, pp. 399–435, 2018.

[48] CFR 16, "Commercial practices–Federal trade commission," in *Part 255–Guides Concerning Use of Endorsements and Testimonials in Advertising*, Oct. 2009, ch. 1. [Online]. Available: https://www.govinfo.gov/app/details/CFR-2020-title16-vol1/CFR-2020-title16-vol1-sec255-0

[49] *Commerce and Trade–Federal Trade Commission*. Washington, DC, USA: U.S. Govt. Publ. Office, 2011. [Online]. Available: https://www.govinfo.gov/app/details/USCODE-2010-title15/USCODE-2010-title15-chap2-subchapI-sec45

[50] M. E. Kaminski. (2018). *The Right to Explanation, Explained*. Accessed: Jul 8, 2019. [Online]. Available: https://papers.ssrn.com/abstract=3196985

[51] *Data Protection by Design and by Default*, Gen. Data Protection Regulat., Eur. Parliament, Council Eur. Union, 2016. [Online]. Available: https://gdpr-info.eu/art-25-gdpr

[52] D. Matthew. *Implementing American Health Care Reform: The Fiduciary Imperative*. [Online]. Available: https://scholar.law.colorado.edu/cgi/viewcontent.cgi?article=1171&context=articles

[53] A. D. Selbst. *Negligence and AI's Human Users, Boston University Law Review, Forthcoming, UCLA School of Law*. Accessed: Mar. 11, 2019. [Online]. Available: https://ssrn.com/abstract=3350508

[54] S. R. Listokin, "Industry self-regulation of consumer data privacy and security," *J. Marshall J. Info. Tech. Privacy Law*, vol. 32, pp. 15–32, 2015.

[55] F. Pasquale and O. Brach, "Federal search commission—Access, fairness, and accountability in the law of search," *Cornell Law Rev.*, vol. 93, pp. 1149–1209, 2008.

[56] J. M. Balkin, "Information fiduciaries and the first amendment," *UC Davis Law Rev.*, vol. 49, pp. 1183–1284, Apr. 2016.

[57] L. M. Khan and D. E. Pozen, "A skeptical view of information fiduciaries," *Harvard Law Rev.*, vol. 133, pp. 497–541, Sep. 2019.

[58] J. Collins, *Turning the Flywheel*. New York, NY, USA: Harper Collins, 2019.

Anthony Aguirre was born in Los Angeles, CA, USA, in 1973. He received the B.S. degree in mathematical physics from Brown University, Providence, RI, USA, in 1995, and the Ph.D. degree in astronomy from Harvard University, Cambridge, MA, USA, in 2000.

He was a Postdoctoral Fellow with the Institute for Advanced Study, Princeton, NJ, USA. He is a Faggin Professor for the Physics of Information with the Department of Physics, University of California at Santa Cruz, Santa Cruz, CA, USA, and part of the Santa Cruz Institute for Particle Physics. While his primary research focus is theoretical physics, information theory, and cosmology, he has co-founded two institutions related to AI and technological innovation. The first is the Future of Life Institute, a nonprofit research, outreach and advocacy organization addressing transformative technologies; it has organized multiple landmark conferences on AI safety and ethics, and coordinated formulation of the Asilomar AI principles. The second is Metaculus, a massive science- and technology-focused forecasting platform.

Gaia Dempsey was born in Rome, Italy, in 1986. She received the Bachelor of Arts degree from New York University, New York, NY, USA, in 2009, with a dual major in urban design and architecture and comparative literature in three languages.

She currently leads 7th Future, Santa Barbara, CA, USA, a consultancy focused on developing resilient, scalable technology infrastructure, and coordinating engagement between educational institutions, private industry, and government stakeholders. Prior to 7th Future, she co-founded DAQRI, an augmented reality hardware company offering a complete professional AR platform to the industrial and enterprise market, as well as Demeter Interactive, a digital marketing and communications firm for tech startups and public–private partnerships focused on sustainability. Sought out as a thought-leader, writer, and speaker on AR, AI, and innovation, she has spoken at Augmented World Expo, IoT World, SXSW, Sundance, the Wearable Technology Show, the Future of Storytelling, Web Summit, and Collision.

Harry Surden received the B.A. degree from Cornell University, Ithaca, NY, USA, in 1995, and the Juris Doctor degree from Stanford University, Stanford, CA, USA, in 2005.

He is a Professor of Law with the University of Colorado Law School, Boulder, CO, USA, and affiliated faculty with the Stanford Center for Legal Informatics (CodeX), Stanford University. Before joining the University of Colorado, he was a Resident Fellow with Stanford Law School, Stanford. Prior to earning his law degree, he worked as a Professional Software Engineer with Cisco Systems, San Jose, CA, USA, and Bloomberg L.P., New York, NY, USA. His research focuses upon legal informatics, artificial intelligence and law (including machine learning and law), legal automation, and issues concerning self-driving/autonomous vehicles. He also researches and writes about intellectual property law with a substantive focus on patents and copyright, and information privacy law.

Peter B. Reiner (Associate Member, IEEE) was born in Nyíregyháza, Hungary in 1955. He received the V.M.D. and Ph.D. degrees from the University of Pennsylvania, Philadelphia, PA, USA, in 1982 and 1984, respectively.

He was a Postdoctoral Fellow with the University of British Columbia (UBC), Vancouver, BC, Canada, and the University of Zurich, Zürich, Switzerland. He is a Professor with the Department of Psychiatry and a member of the Centre for Artificial Intelligence Decision-Making and Action with UBC, and the Founder of the Neuroethics Collective, a virtual think tank of scholars who share an interest in issues of neuroethical import. He began his career as a member of the Kinsmen Laboratory of Neurological Research with UBC, where he was the Inaugural Holder of the Louise Brown Chair in Neuroscience. In 1998, he became the Founder, the President, and the CEO of Active Pass Pharmaceuticals, and in 2007, he co-founded the National Core for Neuroethics.

Prof. Reiner is a member of the Neuroethics Subcommittee of IEEE.