frontiers | Frontiers in Physics

TYPE Review
PUBLISHED 04 January 2023
DOI 10.3389/fphy.2022.1080132

OPEN ACCESS

EDITED BY
William Frere Lawless,
Paine College, United States

REVIEWED BY
Monique Beaudoin,
University of Maryland, College Park, United States
Judith Dijk,
Netherlands Organisation for Applied Scientific Research, Netherlands

*CORRESPONDENCE
Ariel M. Greenberg,
ariel.greenberg@jhuapl.edu

SPECIALTY SECTION
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

RECEIVED 25 October 2022
ACCEPTED 02 December 2022
PUBLISHED 04 January 2023

CITATION
Greenberg AM and Marble JL (2023), Foundational concepts in person-machine teaming.
Front. Phys. 10:1080132.
doi: 10.3389/fphy.2022.1080132

Foundational concepts in person-machine teaming

Ariel M. Greenberg[1]* and Julie L. Marble[2]

[1]Intelligent Systems Center, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States, [2]Institute for Experiential Robotics, Northeastern University, Boston, MA, United States

As we enter an age where the behavior and capabilities of artificial intelligence and autonomous system technologies become ever more sophisticated, cooperation, collaboration, and teaming between people and these machines is rising to the forefront of critical research areas. People engage socially with almost everything with which they interact. However, unlike animals, machines do not share the experiential aspects of sociality. Experiential robotics identifies the need to develop machines that not only learn from their own experience, but can learn from the experience of people in interactions, wherein these experiences are primarily social. In this paper, we argue, therefore, for the need to place experiential considerations in interaction, cooperation, and teaming as the basis of the design and engineering of person-machine teams. We first explore the importance of semantics in driving engineering approaches to robot development. Then, we examine differences in the usage of relevant terms like trust and ethics between engineering and social science approaches to lay out implications for the development of autonomous, experiential systems.

KEYWORDS

autonomous systems, trust in autonomy, AI ethics, human-machine teaming, semantic mapping

1 Introduction

For much of its history, teaming research has focused on teams of people. Yet, as artificial intelligence and autonomous system technologies become more advanced, we enter an age in which it is necessary to consider how people will team, cooperate, and collaborate with intelligent machines, and *vice versa*. As research on person-machine teaming begins to take shape, the prevailing assumption has been that the social interactions occurring within interpersonal teams (and/or teams including non-human animals) can serve as a useful basis for understanding the interactions between persons and machines. However, we argue that that there are essential differences between persons and machines that require special consideration when discussing person-machine teaming.

In particular, there are foundational concepts in interpersonal (person-person) teaming that require translation and adaptation when applied to person-machine teams. These concepts include *Autonomy* (compared to *Automation*), *Trust* (compared to *Reliability*), *Ethics* (compared to *Governance*), and *Teaming* (compared

to *Use of Automation*) and other, similar terms, which do not directly port from teams comprised of people to those including machines.

While of superficially minor distinction, the interpersonal concepts to which these terms refer inform how we interact with, interpret, and evaluate behavior, regardless of whether their extension to machines is performed casually or deliberately. In teaming, essential notions underpinning these concepts, such as *control* and *vulnerability* (compared to *risk* or *uncertainty*) tend to become mischaracterized or fall out of consideration in the course of translation. In the following pages, we will scrutinize each concept across the interpersonal and machine contexts and identify features that warrant additional consideration in translation.

To accomplish this, we will review conceptual issues that have arisen in the translation of the above italicized terms from the interpersonal context to that between people and machines, and the cross-disciplinary roots of the divergent uses for each term. We begin by summarizing an assembly of computational linguistic techniques devised to shed light on the state of discourse around concepts for application to teaming with machines that we refer to collectively as the *Semantic Mapping Pipeline (SMP).* Then, we provide a series of qualitative discussions about topics ready to be run through this quantitative method, beginning with *Teaming and Sociality*, continuing with underpinning notion of *Vulnerability,* and then covering the concepts of *Autonomy*, *Trust*, *Ethics*, and *Teaming.* We conclude with a discussion that summarizes the major arguments presented.

2 Personhood and relationships over speciesism

While it is common for researchers to use the term "human-machine teaming" and to speak of "the human" doing this or that with "the machine," a core tenet of our analysis is that the species of intelligent animal is not the primary feature that distinguishes people from machines. Instead, a more salient difference between the two classes of teammates, in the spirit of Locke, Singer, and Strawson, is that humans are persons, who are able to reciprocally recognize the personhood of other humans, whereas machines are not (yet, if they ever could be) persons, and are not (yet) able to recognize personhood, even if they can discriminate humans from other species [1]. In making this distinction, we seek to highlight the fact that there is something special about those with personhood status (and people, in particular) that makes their dyadic behavior fundamentally distinct from that of their machine counterparts. Indeed, we believe that the direct comparison of living species with technology is a false equivalence. First, it implies that the two classes of teammates may be treated similarly—that the person is just another cog in the system with an input and output interface, ripe for replacement by machines. Second, the use of the term "the human" distances and objectifies the person with clinical detachment, especially when juxtaposed with "the machine," so reducing people to automatons. Thus, the use of the term "the human" gives the impression that person-person interactions (and the language that is used to describe them) are directly analogous to person-machine interactions, when it is eminently clear that much of what imbues these actions with their significance stems from mental capacities that no machines currently possess. The mutual relationship of personhood and person recognition in the interpersonal sphere is perhaps the absent core that prevents direct translation to person-machine teaming context (more on this in upcoming paper on relationships by Hutler and Greenberg, forthcoming). For the rest of this paper, in deference to the taxonomic superiority of *person* over *human* (a *human* is a type of *person,* and all humans are people), we will try to correct this terminology by referring strictly to *persons* or *human beings* where *humans* might typically be used, in particular to recast human-machine teaming as person-machine teaming or PMT.

3 Semantics matter

When a concept is ported from one domain to another, a typical first step in the engineering design process is for practitioners to compress the concept into an operational definition toward which they can build. This reduction in practice is ordinarily very effective [2]. However, when designing for PMT in particular, salient features of the concept to be replicated (e.g., sensitivity to vulnerability, recognition of personhood) are frequently lost in translation, while skeuomorphic features (i.e., machine features that superficially emulate interpersonal capabilities, but are substantively dissimilar under the hood, e.g., voice production or eye-contact) are unintentionally retained, or picked up in translation, leading to inappropriate expectations of machine capability. In contrast to these issues in the translation of interpersonal capabilities, translation of physical capabilities (e.g., walking or grasping) is relatively mechanistic and straightforward.

There are perils in this lossy compression (The metaphor of *lossy compression* is used here to indicate that the concepts coming out of this processes are smaller, but also lower resolution). If we only use these concepts in their most superficial form, we miss out on the richness to be had in the phenomena they signify. If we use them in their full interpersonal sense, we misrepresent the capabilities of the machine and set inappropriate expectations for their performance (see wishful mnemonics [3]). If we use the terms in an ambiguous sense between the most superficial and the full interpersonal, then the capabilities realized in different machines are bound to be

inconsistent across implementations. This last case is the most prevalent, and with each occurrence, the conceptual drift continues. By allowing this inconsistent usage to prevail, we may ultimately lose our grip on the original meaning, our appreciation of the fullness of the phenomenon may diminish, and the reduced definition may become prone to be cast over the entire phenomenon, interpersonal and otherwise. As Sherry Turkle [4] puts it: "When we see children and the elderly exchanging tenderness with robotic pets, the most important question is . . . *what will "loving" come to mean?*" [emphasis added].

This disconnect in language becomes especially apparent in design meetings and program reviews, wherein operational definitions are only found to be incongruent with empirical capabilities after the fact. Worse is when that incongruence remains unrecognized—the same words are used by different designers with significantly different meanings. This disconnect naturally arises from the different backgrounds of those using the terms. Robotics is inherently a multidisciplinary area of research, and different disciplines understand and use the same terms very differently, including how to measure them in context. Of course, harmonizing terms is a perennial challenge in multidisciplinary work, but is particular acute for social robotics since interpersonal terms have previously been used in the context of technology more metaphorically than anthropomorphically, or simply for purpose of usability. An engineer using the term "trust" may construe the term with respect to things that can be engineered, while a social scientist might construe the term with respect to social constructs. We argue that adequate definitions are those which may be operationalized sufficiently for design and can be measured accurately, reliably, and repeatably, while respecting the richness of the phenomenon in the context of its relevant interpersonal constructs.

There are a number of interventions possible to address this disconnect. The most extreme is to declare that interpersonal terms may not be used in the context of person-machine teaming. This prescriptive approach to controlling language rarely succeeds, and will only be effective at alienating incoming generations of researchers. Another intervention is to create new terms or to add a qualifying adjective to existing ones to make these terms specific to PMT [e.g., adding *semi*-to qualify *Semi*-autonomous [5], adding *robot and artificial intelligence* to qualify *RAI*-responsibility (upcoming publication by Greenberg et al., Robots that "Do No Harm")]. The most gentle intervention is to do what we have set out to do here: Identify what is lost and found in translation between contexts.

3.1 Semantic mapping pipeline (SMP)

As part of an earlier effort, Greenberg led a small team to review how terms central to person-machine teaming were being used across the literature. This preliminary investigation sought to develop the methodology and begin a cursory exploration, and it is presented here to introduce a mixed-method approach to semantic conceptual analysis. In the sections that follow, we discuss the concepts of *Autonomy*, *Trust*, *Ethics*, and *Teaming* in primarily qualitative terms. However, we believe that these same topics are ripe to be run through the SMP method for quantitative support.

This review, formulated as a semantic map of terms, was intended to address questions such as:

- How do various disciplines use PMT terms, both within their discipline, and when communicating to their interdisciplinary counterparts?
- What are the differences and similarities in the ways the various disciplines use the terms?
- How much true interdisciplinary treatment is there, or is treatment mostly disparate multidisciplinary contributions?
- In which semantic clusters does a particular organization find their conceptualization to fit best (the inverse problem of semantic map assembly, whereby particular articles invoking the terms are placed within the map generated from the corpus).

To answer these questions, the concepts under consideration are first cataloged and discussed. Once a suitable list of keywords has been identified, they are then run through the semantic mapping pipeline to display their prevalence, authorial provenance, and co-occurrence in current person-machine teaming scholarship, against a background of those terms' usage in interpersonal contexts and in common parlance. The methodology of the semantic mapping pipeline is as follows:

First we populate a corpus; the body of papers that include the terms of interest, semantically similar terms, and their related word forms across various parts of speech, retrieved from scholarly clearinghouse sources like Web of Science and arXiv, and from policy statements of international organizations. We then review the bibliographies of these papers to augment the corpus with secondary papers that are related but may not have used the search terms precisely as we had specified them. Given that the contents of this corpus is the source material from which the pipeline produces it analyses, we take care to be comprehensive at the start. Late additions are possible to be accepted, at which point those new entries are reprocessed as described in the next steps, for an updated output.

Next, we use the systematic review software Covidence to screen the papers for relevance by the PRIMSA protocol, and tag them with an interpretation of how the paper authors are using the search term, from a standardized list of meanings set *a priori* from a preliminary scan. Should new meanings be discovered in-process, they are added to this list for tagging. Those papers emerging from the screen are parsed, along with their metadata.

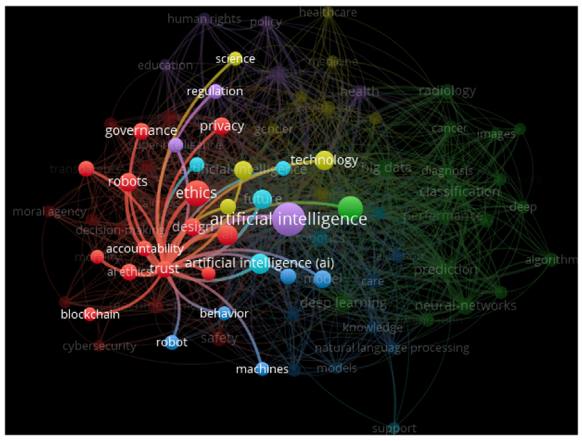

FIGURE 1
Term co-occurrence.

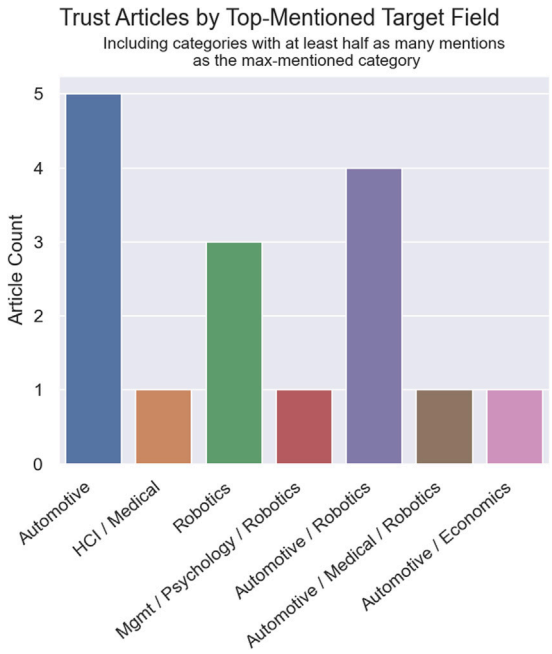

FIGURE 2
Articles by target.

With the corpus now formed into a computational object, we can visualize and analyze *in silico*. For preliminary visualization, nodes of terms are linked by edges of co-occurrence, sized by number of citations, and colored by community detection that reflects disciplinary field. A semantic graph composed of these nodes is assembled and visualized by VOS Viewer (Figure 1), or programmatically by the python Network X package.

Next, we perform various analyses to update the semantic map. Bibliometric analyses include undirected graphs of cocitations and directed graphs describing the discourse between contributing disciplines, authorial provenance, and target audience (Figure 2). Throughout, the method invokes *synsets* (WordNet's grouping of synonymous words that express the same concept) to improve flexibility across semantically similar terms. Semantic analyses encompass the usage patterns of terms found across several parts of speech: nouns, adjectives, and prepositions.

The nominal [of nouns] analysis queries co-occurrence graphs and compares term frequency distributions (Figure 3) to discover if PMT discussions around a term are addressing the same concept—and if so, how those discussions are distinct by discipline and/or from usage in interpersonal literature. The adjectival analysis queries the lexical dispersion (relative locations of terms within the text and their distances from one another, Figure 4) of preidentified terms and of terms with wordfinal morphemes of adjective suffixes, to collect how terms are described, and whether descriptions are used consistently throughout each document. The prepositional analysis uses phrase chunking to discover to what, to whom, or for what the term pertains.

By filtering the visualization of the semantic map that serves as frontend to the combination of computational linguistics techniques here described, researchers may examine term semantics at scale. We invite those interested to adopt this mixed-method approach and continue the work where we left off, with code available upon request.

4 PMT concept 1: Teaming is inherently social

Unlike almost any other engineering product, autonomous systems interact with people through social channels to achieve their goals. Meanwhile, people's responses even to non-agentic computers are inherently social (e.g. [6]), and with just a bit more interactivity, they become what Sherry Turkle [4] calls relational artifacts: "Their ability to inspire relationship is not based on their intelligence or consciousness but on their ability to push certain Darwinian buttons in people (making eye contact, for example) that make people respond as though they were in a relationship."

The question of whether a machine can truly team with people (or even other non-human agents) is a source of significant debate, and the term "team" is frequently misused or misapplied, especially with respect to person-machine teams. Research engineers often apply the term "human-machine team" to any collection of people and robots or autonomous agents, regardless of whether they meet the criteria that define a team, such as the need for interdependence between members or common identity as a team [7–9].

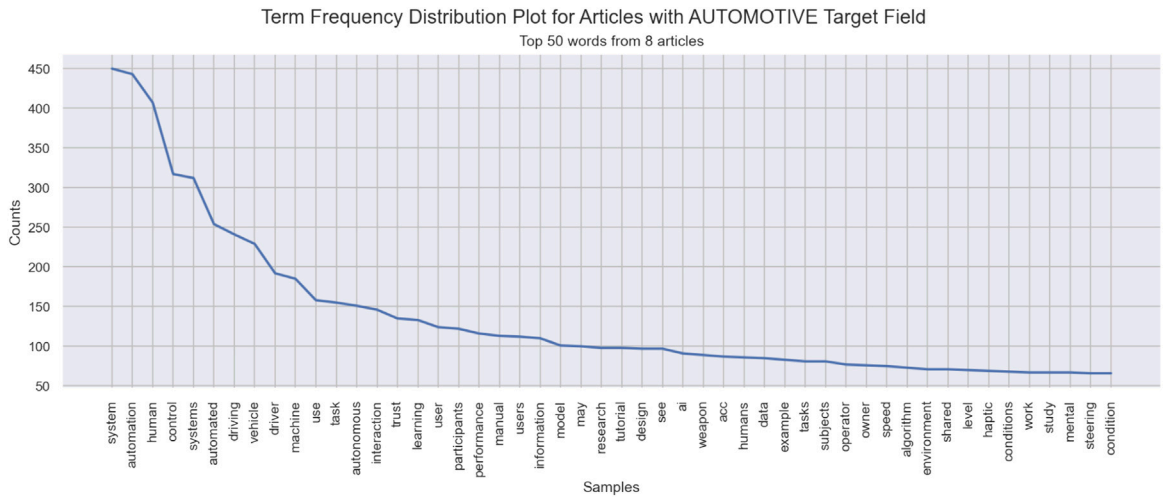

FIGURE 3
Term frequency distribution.

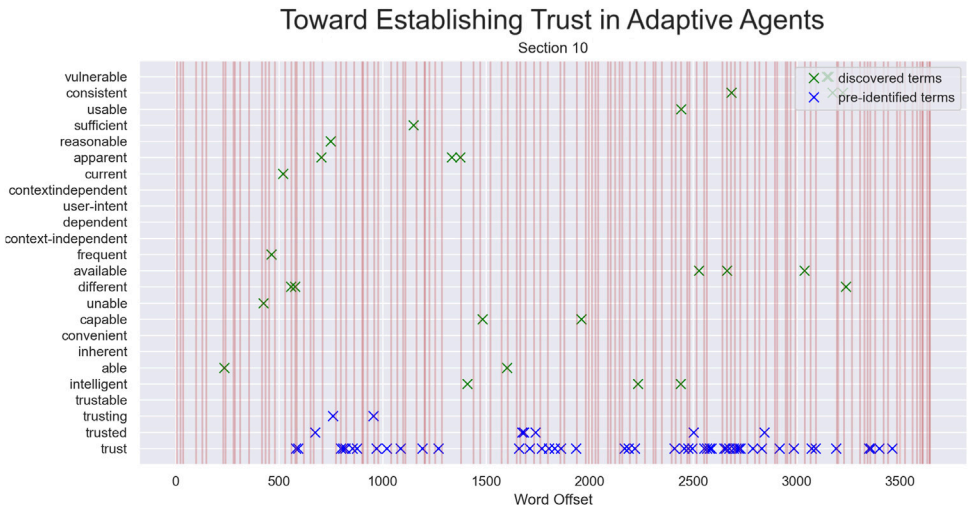

FIGURE 4
Lexical Dispersion, with sentence breaks indicated by red vertical lines. The legend is on top of hits for the first two discovered terms ("vulnerable" and "consistent") between word offsets 3000 and 3500.

A team is a set of two or more people who interact dynamically, interdependently, and adaptively, toward a common and valued goal, each member having specific roles or functions to perform, and a limited life-span of membership [10]. Teams, therefore, are inherently social groups with interdependence between team members who are working toward common goals [10–12]. Team members behave differently from other organizational structures (e.g., supervisory hierarchies) in several ways. They demonstrate increased communication between members, greater effort and commitment to the goals, greater trust between members [13] and show greater adaptability and innovation from these other structures [14]. Kozlowski and Ilgen [14] also emphasize the social aspects of teaming—motivation, affect and interpersonal interaction.

As a further example, Walliser et al. [11] explored how team structure and the manner in which people are directed to work with a teammate impact team performance with autonomous agents. In this study, participants worked with a human collaborator or an autonomous system, either as a

collaborative teammate or to direct its performance as they would a tool. As would be expected, collaboration was more common in the teaming condition than in the tool condition. For example, there were significantly more task-relevant chat messages sent by participants in the team condition. Task-relevant chat messages were equally common for both the human and autonomous agents. In contrast, messages related to performance, information, and acknowledgment were only sent when the other agent was human. The authors argue that these results indicate that the interaction between people and autonomous agents is fundamentally social; given that effective teamwork relies heavily on social interactions, these aspects of interaction must be included in the development of autonomous agents. They point out that the social aspects of person–machine team design are neglected in favor of enhancing the more traditional computational and electromechanical capabilities of the autonomous agent. We explore that focus in the next section where we examine guidance given on the design and development of autonomous systems.

The debate regarding whether autonomous machines may be considered teammates over tools centers on the development and demonstration of shared common goals or shared mental models, interdependence of actions, and inter-agent trust [15]. Relatively recent advancements have begun to demonstrate the ability for machines to share goals and adapt to changing context (see for example [16]). Further, people appear to team as easily with robots as with humans [17, 18]. Taken together, these findings suggest that research that neglects the experiential, social, and cognitive-affective aspects of person-machine interaction will not yield successful teaming; in which case machines will remain in the role of tools and the full capabilities of effective person-machine teams will not be realized.

One way to approach these neglected aspects of PMT is to attend to the latent construct of vulnerability. The constituent concepts we will review in the next sections, on Autonomy, Trust, and Ethics, all share this latent construct, which tends to be the first to fall out when translating these terms from their interpersonal sense to their person-machine teaming sense. *Vulnerability*, the state in which a person is subject to harm (physical, psychological, financial, etc.) remains the condition for a teammate whether that person is relying on another person, or on a machine.

5 PMT concept 2: Vulnerability is ultimately unmitigable

In PMT contexts, the notions of autonomy, ethics, and trust are inextricably linked not just to mission and task risk (cognitive trust [19]) but to personal vulnerability (emotional trust, [19]. To demonstrate this for yourself, try the following exercise—replace the terms *autonomy*, *ethics*, and *trust* with a conjugate of *vulnerability*, and determine whether the statement still holds[1]. However, while this connection is apparent in every definition of interpersonal trust (see [20, 21]), the notion of vulnerability is frequently operationalized as relatively less-rich concepts such as uncertainty or risk when translated to pertain to persons cooperating with machines. This may be because *vulnerability* is perceived as more affect-laden and nebulous, while *uncertainty* or *risk* can be defined in probabilistic terms, which is more compatible with an engineering orientation. However, the notion of vulnerability is not encompassed by uncertainty or risk alone, and creating an operational definition that exchanges these concepts loses essence (now try that term replacement exercise again, with *uncertainty* or *risk* swapped for *vulnerability*). The stakes are not simply outcome- or likelihood- oriented pertaining to risk, but indeed personal—a machine teammate's failure has personal consequences for its human teammates.

These consequences may arise not just from failure to complete the task (as discussed in Section 8 on trust), but from performing the task in unexpected or incompatible ways, or from performing the task in an expected manner that yields undesired results. Among other things, human teammates may grow disappointed, insecure, or worried, and that negative affect is itself a harm, not captured by the concept of risk (though approximated by *vigilance*). While this may not appear to be a consequential effect, keep in mind how crucial a lever negative affect is for humans teaming with non-human animals: dogs in particular are exquisitely sensitive and responsive to our disposition to them [22].

Of course, the typically negative affect associated with the experience of vulnerability is not felt by machines, so there is an intrinsic limit to how faithfully a machine can participate in the downstream concepts of PMT Autonomy, Trust, Ethics, and Teaming. As put by Marisa Tschopp [23]: "The victims are always the humans." Even just an imbalance of vulnerability between partners is generally enough to undermine trust [24]. Autonomous systems are indifferent about survival; are without social or emotional values to protect; are unconcerned with stakes and unaffected by reward (despite it being sought computationally through reward and objective functions in machine learning) and undeterred by punishment. Autonomous systems have nothing to lose, and nothing to gain, so the act of judgement must be privileged to those who are innately vulnerable (people), who also have a sense of responsibility and who are affected by the potential disappointment of those subject to the judgement.

1 For example, does "I *trust* the machine to fold my laundry" mean the same as "I *am willing to be vulnerable to a poor outcome* should the machine not succeed," or simply that "I *believe* the machine will be successful?"

Further, machines do not have the visceral appreciation for human vulnerabilities that people do. As a result, people have no basis for confidence that machine teammates will understand the shape of the utility functions of people to select a behavior that is congruent with their interests. This creates inter-dyadic risk that is independent of the operational context (or, at the very least, omnipresent across all contexts), and dramatically lowers the likelihood that people will be willing to trust the machine. It is not just that machines do not share the same vulnerabilities, it is that because they cannot feel vulnerable, we don't expect them to share or understand our values.

To address this vulnerability gap, Greenberg has worked toward the development of a harms ontology, described further in the section on *Ethics*. In this installment of research into artificial non-maleficence, he and his team explicitly trace potential physical harms to humans through their vulnerabilities (in this one case, the biology of the species of intelligent animal is the salient feature vs. their personhood that is primary for the other ethical principles and types of non-physical harms). From an ethical standpoint, each actor should seek to recognize and respect the vulnerability of other actors, to minimize harms that prey upon that vulnerability. In fact, the ability to recognize vulnerability may be a criterion for personhood (Strawson Microsoft Word - Document3 (brandeis.edu)).

In both interpersonal and PMT contexts, *Control* is the primary means to mitigate vulnerability to another actor's behavior or to situation outcomes. It is also something engineers are adept at building the means to achieve (e.g., control theory, control surfaces, controllers, etc.). However, increased control by people of machine actions diminishes the machine's independence, defeats the objective of autonomy, and squarely eliminates the opportunity for trust, which otherwise thrives when the trustor's vulnerability is protected by the trustee amidst unpredictable circumstances, even if (or especially when) the objective may not be met, but the measures to protect are communicated and appreciated.

McDermott et al. [25] provide an example of vulnerability mitigation in their guide on development of human-machine teaming systems [the authors of this guide use the term *human*—in the context of this paper, we would use the term *person*]. In their guide, they first discuss "Directability." Directability is supported when humans are able to easily direct and redirect an automated partner's resources, activities, and priorities. People will have expertise or insights that the automated system will not. People are ultimately accountable for system performance, which means they must be able to stop processes, change course, toggle between levels of autonomy, or override and manually control automation when necessary. They provide the following guidelines for development:

- The automation/autonomy shall provide the means for the operator to redirect the agent's resources when the situation changes or the human has information/expertise that is beyond the bounds of the agent's algorithms.
- The automation/autonomy shall provide the operator the capability to override the automation and assume partial or full manual control of the system to achieve operational goal states.
- The automation/autonomy shall not remove the human operator from the command role.

Despite the essentialness of vulnerability to PMT concepts, the term is rarely operationalized in any meaningful fashion within discourse or experimentation. In scanning our SMP corpus for lexical dispersion of the term, we find that it frequently appears in isolated statements and definitions, but is otherwise abandoned [20]. In fact, the interpersonal definition of vulnerability is often contramanded by experimental design. In the excellent review by Woolley and Glikson on Trust in AI, the authors open with Mayer's definition of interpersonal trust: "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [26]. In contrast and in contradiction, Woolley and Glikson's summary of research conducted by Ullman and Malle states: "They found that participants reported higher cognitive trust in the robot *they controlled*" [emphasis added]. Furthermore, following this controlled experience of involvement, participants expressed significantly higher trust in potential future robots [27, 28]. This discrepancy is apparent but not addressed: If trust is about willingness to be vulnerable irrespective of control, then what is an experiment truly measuring if it finds that "trust" is contingent on level of control? Further, trust entails an acceptance of vulnerability, which is refused by a desire to control.

Recently, the authors of this paper explored the development of trust between people and machines in using a virtual environment, the Platform for Assessing Risk and Trust with Non-exclusively Economic Relationships (PARTNER) [We refer to this experiment to illustrate the questions of interest rather than to elucidate the results, therefore, we will forgo review of the conclusions. Interested readers can refer to [17]]. In PARTNER, people and machines are paired up to escape a room, and these puzzle stages are constructed to be unsolvable without cooperation (see Valve's Portal 2 game). We made sure to draw in and probe vulnerability in two ways: The first was to build upon its operationalization in Berg et al.'s [29] canonical investment game concerned with financial trust. We argued that the paradigm used—to give a gift of funds which may then be lost during the interaction—did not invoke authentic vulnerability in most participants; thus we focused on non-economic relationships. We argued that inducing participants to experience a sense of physical vulnerability comparable to a trust fall would be more

effective and relevant in the real world [A trust fall is a team-building exercise in which a person deliberately falls, trusting the members of a group (spotters) to catch them.]. Insofar as Institutional Review Boards (IRBs) generally frown upon the prospect of dangling people off the edge of cliffs, we opted to do so in virtual reality (VR), from heights and into pits of hazards, to emulate physical peril. Falling in VR is a reliable method to trigger the sensation of falling in the vestibular system, and some users even experience vertigo (those participants were screened out). The other way we enabled opportunities to experience non-financial vulnerability was by creating situations for the robot partner to save or betray the human player, and for the robot partner to perform activities that were hazardous to the human (again, those hazards relate in particular to human biology). Which teammate took the risk-laden action exposes an aspect of trust designed for experimental examination: Did the person perform the safe task while the robot took the risky task (e.g., the task with the potential to fall)?

6 PMT concept 3: Autonomy is a relationship, not a system property

The term autonomous systems (AS) has its origin in warfare. The person-machine unit of a submarine is the typical exemplar, often separated from traditional C3 (command, control, and communications), and authorized to act without instruction. A special class of autonomous systems, lethal autonomous weapons systems (LAWS), are machines set to fire when conditions are met in cases in which intervention by people would be too slow to neutralize the threat. When LAWS are referred to as human-machine teams, the macabre reading is that people are participating only in the sense that they are the targets. LAWS do have significant bounds and limitations on their behavior: The systems cannot not act if the conditions for action are not met, nor can the systems weigh factors in the environment which have not been programmed to assess. While these machines are able to perform complicated actions without the direction of a person, in many respects LAWS are still more automated than autonomous[2].

When used in an interpersonal context, the term autonomy is meant to indicate that a person is not subject to another authority in making personal determinations. This sense of autonomy concerned with self-governance is not even desirable for installation in machines—after all, autonomous systems are meant to improve the human condition and serve people's needs, not act as machines want for themselves (as if wants are even possible for machines). Autonomous systems are artificial and designed, and thus without true motivations. In contrast to automation, wherein a technology performs a pre-specified task in a controlled environment, machine autonomy (in the PMT context) is often used to describe sophisticated, flexible, or adaptive automation that can perform with some degree of initiative and independence in novel contexts or environments, without complete external oversight or control. Importantly, autonomy is earned and awarded through an external authority, making it a property of a relationship rather than a property of an entity within that relationship, as in automation.

Autonomous systems are commonly understood as decision-making technology both capable and worthy of being granted some degree of independence from human control. However, "decision-making" as used here is wishful mnemonic (cf. [3]) for the calculations these machines perform, and the actuations to accomplish the determination of those calculations. While the systems do hold goals, objectives, and missions, these imperatives exist around the level of programing. These systems do not really make decisions, conduct judgements about the preferability of different actions, or emergently generate novel options to choose amongst beyond the methods available in their deployed code.

Currently, potential options and actions available to autonomous systems are limited by their programming, but these machines may eventually be so capable that available to them are such a broad spectrum of possibilities that the limits to their actions cannot be fully predicted; in fact, in systems that are not embodied in the physical world, such as on-line avatars or large language models[3] we are rapidly approaching this uncircumscribed scope, if we have not already reached it.

Though autonomous capability and intelligence often overlap, they are distinguishable. Where autonomous capability is concerned with initiative and independence, intelligence is concerned with the ability to hedge against dynamic vulnerabilities—i.e., threats to autonomy, coordination (teamwork), and ethical (desirable) behavior—in real time. In other words, intelligence and agency are among the essential components of the "personhood" that's missing. For a study in the topic of intelligence, see an upcoming paper in *Entropy* by Baker and Greenberg.

Machine autonomy is not a widget that can be built [30], but rather a privilege people grant to machines that are capable of operating without or outside of our supervision and control. That privilege is earned after testing and experience have demonstrated the capability, or in cases where control is impossible due to environmental constraints (remote, dirty, dangerous). Various conceptual efforts [31, 32] to arrange autonomy as levels, as adjustable, or on a sliding scale, falter

2 For further information about C3 and LAWS, please see these references: Chapter 20 Command, Control, and Communication (fas.org), IF11150 (congress.gov), DoDD 3000.09, 21 November 2012, Incorporating Change 1 on 8 May 2017 (whs.mil).

3 Is LaMDA Sentient? — an Interview | by Blake Lemoine | Medium: Though the authors of this paper do not accept the sentience claim, the novelty claim is compelling.

in ordering autonomy as a single functional unit, as opposed to a collection of constituent capabilities that combine in complex patterns to enable minimal communication along the appropriate level of abstraction. These constituent capabilities included in the notion of autonomy, initiative, and independence, and in particular, graceful handoff, are buildable.

Ideally, we might want people to be the ones drawing the line for transfer of attention, but in practice, it may have to be determined by machines, driven by time constraints to be part of its autonomous functionality. Accomplishing effective and efficient handoff between machines and people requires substantial social cognition on the part of the machine. First, the person-machine system needs to assess whether an action is in the purview or even the ability of the machine or the person. Not only does the machine need to know its performance boundaries, that is, what it can and cannot do well, but both the machine and the person require the bit of metacognition that allows each to infer what the other does not or cannot know or do. Together, these indicate to the machine when it ought to ask for help from people, for the person to offer assistance, or that it is not appropriate to ask for assistance. If the machine determines that it cannot or does not know information critical to performing the task, or that it does not have the capability to act, it needs to ask for help. In that respect, autonomous systems should be experiential—they should learn from their interactions with people, or from the experiences of other autonomous systems. Critically, methods are needed to ensure this learning is indeed in the desired direction, and that the autonomous system will not converge to performance boundaries that are unwanted. Appropriate requests for assistance require that the machine have elementary theory of mind, that is, to infer who might know what, who to ask, and deixis (how to refer in time, space, and person). Finally, the machine may need to escalate the request for attention to a person, and hand off the question or task to them. Requests for assistance cannot happen all the time or the system is almost useless, nor can they never happen as the system would take unacceptable action or fail to act appropriately too often. Similarly, if the task is to be handed off, there must be sufficient time for the person to assess the context and prepare to perform the task, as well as to perform the task (Tesla Autopilot Crashes into Motorcycle Riders—Why?[4] 7:24: "So before you trust his take on autonomy, just know that autopilot is programmed to shut down one second before impact, so who's the manslaughter charge going to stick to?"). The timing, information provided, and receptivity of the person are elements of this handoff package. The machine should not escalate for attention matters that set up the people for failure, by leaving insufficient time or providing insufficient information for the issue to be adjudicated by people, or by sharing with people who are not available to receive the handoff. This means that developers must consider the full spectrum of activities in which the person might be engaged as it is a person-machine team wherein neither entity is fully separable.

7 PMT concept 4: Ethics for machine teammates

Ethics for autonomous systems (i.e., those that make for machine teammates), differ from the ethics of artificial intelligence: In particular, the autonomous system's special features of agency, physicality, and sociality, draw in considerations beyond those of traditional technology ethics concerned with social implications of the built world, to include among other specializations, philosophy of action and philosophy of mind.

Agency, the capacity of an entity (agent) to "instantiate intentional mental states capable of performing action," is not necessarily required for a machine to be granted some degree of autonomy, but that capacity becomes increasingly relevant as these machines are permitted entry into more complex environments. Here, complexity is not strictly along the physical or computational dimensions, but the social—arguably, the environment of a home healthcare aid robot is more complex than that of an autonomous vehicle. *Moral agency*, wherein "an agent has the capacities for making free choices, deliberating about what one ought to do, and understanding and applying moral rules correctly in the paradigm cases" is a much higher bar. It is not clear that machines will ever be able to meet these criteria [33] or even need to in order to accomplish their directives, but *Ethical Agency* is within reach and essential for appropriate system performance. The related concept of *Moral Patiency*, the capacity to feel pain or pleasure remains in the realm of living creatures, and respecting that capacity is the mandate of artificial ethical agency.

Physicality: Not all AI is embodied, and not all autonomous systems can be deemed to have intelligence (not even all AI can be deemed to truly have intelligence, cf. upcoming Baker and Greenberg paper). The ethical implications of a machine with the ability to sense an object in the environment to change direction and avoid it differ from those of algorithms that can crunch large amounts of data. Artificial autonomous capability is generally embodied in a cyber-physical system, and is bound to have direct and indirect effect on the physical world. This is not necessarily true of AI, in which its effects on the physical world tend to be mediated by its provision of information to people.

Sociality: The ethical concerns around machine teammates tend to fall more around *how* the team interacts, how the handoff between team members is performed, and whether each member is prepared to act and capable of acting. If a machine is working with a person, it must perform the handoff between tasks,

4 https://www.youtube.com/watch?reload=9&v=yRdzIs4FJJg

information, objects in such a way that the person is capable of succeeding, while at the same time ensuring that potential for harm to people is minimized. There will be times when a machine is performing a task, and the context changes such that the machine is no longer able to perform the task safely. In those instances, the current development approach is to hand the task back to the person. But this does not ensure that the person is able to perform the task either. It assumes that the person is fully engaged in the task to the point that a hand off is possible. But the purpose of autonomy is to allow the machine to perform without the person, enabling the person to be engaged in other tasks. If the machine is unable to perform the current task, it may be better to have the machine alert the person and instead perform a task at which it is capable of succeeding. In handoff, simply assuming that the person is ready to perform yields a liability issue, and may defy the concept of operations for which the system was built. Rather than transparency of decision making, the person needs to accurately understand how context and environment may affect the ability of the machine to perform. Similarly, the machine needs to understand how the task and environment may have impacted the person's ability to perform, e.g., whether the person has sufficient time to engage in the task, can sense the data or object that has confused the autonomous system, or is even available to perform the task.

Ethics with respect to people refers singularly to the *moral principles that govern a person's behavior or the conducting of an activity*. These principles collect as the set: Transparency, Justice and fairness, Non-maleficence, Responsibility, Privacy, Beneficence, Freedom and autonomy, Trust, Sustainability, Dignity, Solidarity. However, ethics with respect to machines carries at least two senses [34].

The first sense (the *Ethics of machines* or machines as objects of ethical consideration) is the one commonly understood when invoking the terms AI Ethics, or Ethical AI, concerned with the ethical use of artificial agency. This sense in the vein of technology ethics governs human beings (and their institutions), in producing or interacting with machines (their design, use, or interpretation of machine products). The constraints in such governance is extrinsic to the machine, and ethical principles pertain to designers and users. Of the set of principles, fairness, bias, and privacy most exemplify this *of/as objects* sense. Policy documents are exclusively of this sense, both those prescriptive, like from the US Government (IC, DOD, and CIV), Asilomar, and from the Vatican, as well as those descriptive, like the reports by Harvard and Montreal.

The best of breed survey of the *of/as object* sense is Jobin et al. [35]. In reviewing the landscape of AI ethics they came to a consensus around the set of principles listed upfront. They also identified four divergences in how each principle was addressed in the corpus they examined: how ethical principles are

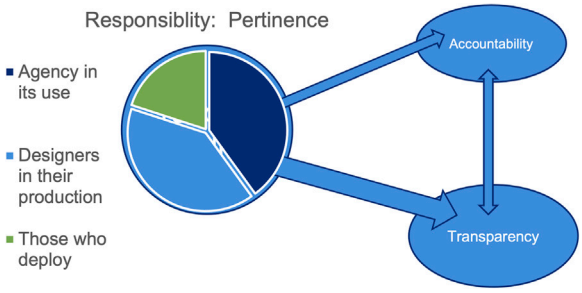

FIGURE 5
Divergence graph.

interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented. These divergences characterize the splay in semantics mentioned earlier.

As part of the SMP effort, we sought to computationally represent and visualize these divergences. In Figure 5 below, we depict a "divergence graph" for the principle of responsibility. These graphs show how different usages or senses of terms (corresponding to Jobin's divergences) differentially connect to related terms. Nodes are sized by term prevalence in the document or corpus. Edges are directional and sized by co-occurrence so that, for example, the width of the link from *responsibility* to *accountability* is to be understood as proportionate to the number of mentions of *accountability* in discussions of *responsibility*. Within the node of *responsibility*, the pie chart indicates the proportions of pertinence usage (Jobin et al's divergence regarding to what or to whom the principle pertains), answering the question: *As it appears in documentation, in what proportions does responsibility pertain to the agency* (in this case, meaning institute) *in its use of AI, designers in their production of AI, or to those who deploy AI?*

The works Jobin review are focused on the ethical *implications of* AI and how policy and governance should safeguard development and protect users. Although this research concerned with obviating and mitigating the personal and societal consequences of AI, such as those presented by algorithmic bias and reward hacking, is crucially important to undertake, it is not the whole picture.

The alternative sense, Ethics *for machines*, or machines as subjects (*for/as subjects*), is the more Asimovian [36] sense concerned with Artificial Ethical Agency. Ethics in this sense regulate the machines themselves, and are only applicable to machines that possess the capability for autonomous agency, unlike other powerful technologies without such a capacity for initiative (like nukes). Ethics for machines are on-board the system proper, and the principles are intended to pertain to the artificial agent itself. This sense of ethics requires commensurate capability and judgement from the machine, a tall order since machines are

ordinarily produced for capability, leaving the judgement for people. That gap is how accidents of the kind at the Moscow Open can occur, in which a chess-playing robot broke a child's finger (for a discussion of this incident see upcoming Elsevier chapter by Greenberg on enabling machines to reason about potential harms to humans). Of the principles, non-maleficence and beneficence are the most clearly of this sort. Important questions about how to "teach" ethics to machines emerge of this sense (described in upcoming robots that do no harm paper). The best of breed survey of the for/subject sense is by Tolmejer et al. [37].

When these two senses are set for and followed by people, there is a unitary apparatus for producing, understanding, and executing the principles. However in machines, these two senses are differentiable, though the *of/as object* sense tends to dominate. To see how little these two senses conceptually overlap between Jobin and Tolmejer, see Figure 6 below.

We argue that successful application of ethics to autonomous systems is distinguished by its goal to explicitly design into machines the basic mental faculties (including perception, knowledge representation, and social cognition) that enable them to act as ethical agents. These capabilities in ethics *for* artificial agents are so fundamental, treatment of them tends to be neglected, but it is at this low and early level that the machine's agency is most available to adjustment by normative considerations. Furthermore, owing to these faculties' universality across major schools of philosophical thought (deontological, consequentialist, and virtue ethics) their essentialness is fairly uncontroversial. Beyond this basic level where mental faculties enable machines to have consideration for moral patients, application of ethics to machines begins to resemble the ethics *of/as objects* sense, wherein appropriate behavior is imposed by governance, leading to brittle performance and diminishment of the capacity for trustworthy autonomous activity.

8 PMT concept 5: Trust is learned, trustworthiness is earned

Trust is a socio-affective construct indicating the willingness of a person to be vulnerable to the unpredictable actions of another. Of the foundational concepts for translation from the interpersonal context to the PMT context, confusion around the term *trust* is perhaps the longest lived and most fraught. The topic of *trust* is also the most integrative of the foundational concepts in PMT, and for this reason we discuss it last. As compared to the rich interpersonal concept, its use in machine contexts is austere. Notable contributions to distinguish the senses between contexts include Thin vs. Thick Trust [38], and Cognitive vs. Emotional [19]. In this section, we first survey the features of interpersonal trust and the intricacies of instantiating them in machines, to then we address issues in measurement and in calibration.

When held between people, trust and trustworthiness are understood to be part of a relationship wherein three characteristics of the trustee make it so that the trustor may confidently hold the belief that the trustee will act in the trustor's interest: ability (or competence), benevolence, and integrity [26]. The stability of one's trust varies depending on which of the aforementioned qualities it is based. If trust is based (solely) on the ability of a trustee, trust should then vary depending on how well the trustee performs a task. If trust is grounded in the integrity of a trustee, then it should vary based not on the actual performance of a trustee but on the extent to which the trustee's actions match the values of the trustee. The stability of benevolence-based trust is contingent upon whether the trustee's actions match the goals and motivations of the trustee. When trust is based primarily on the integrity or benevolence of a trustee, poor performance alone will not significantly damage it. Machines, however cannot truly be either benevolent or malevolent, or have integrity or be corrupt. Researchers have attempted to translate benevolence [39] and integrity for machine contexts, but since these qualities are currently impossible to instantiate in machines as they appear in people, they must be inherited by machines from the people who design them. When the trustee is a machine, the final pillar of trustworthiness—ability—is reduced to little more than "predictable performance," or reliability. This hollow port begs the question of why we bother with this artifice of "Trustworthiness" at all.

Yet researchers continue to pursue designs for autonomous systems that are inherently *trustworthy*. From an engineering perspective, one way to operationalize *trustworthiness* is to ensure that the behavior of the machine is reliable to a high degree, and that the machine is capable of performing the task of interest or telling the person that it is unable to perform the task. From a psychological perspective, based on research on the development of trust between people team members, research demonstrates that these are not the critical bases of the development of trust between team members.

Reducing *ability* to *reliability* is problematic: creating machines that are 99.9% reliable may actually be detrimental to the development of trust in autonomous systems. *Reliability* is defined as the consistent performance of an action, an attribute of the trustee, while trust is a learned response by the trustor [40] applicable to situations in which the trustee is not perfectly reliable, or in which the task entrusted is not certainly achievable. We know from research on learning that consistent reinforcement of behavior does lead to learned response. However, if a consistent reward is discontinued, the learned behavior is quickly extinguished. In other words, if a system is 99.9% reliable, then 999 times out of 1000, it will behave as expected—yielding the learned response of trust. But on that 1000th trial, in response to a system failure, the person's learned response can be quickly extinguished. Variable reinforcement, by contrast, leads to acceptance of a much longer duration without reinforcement before the learned response is extinguished.

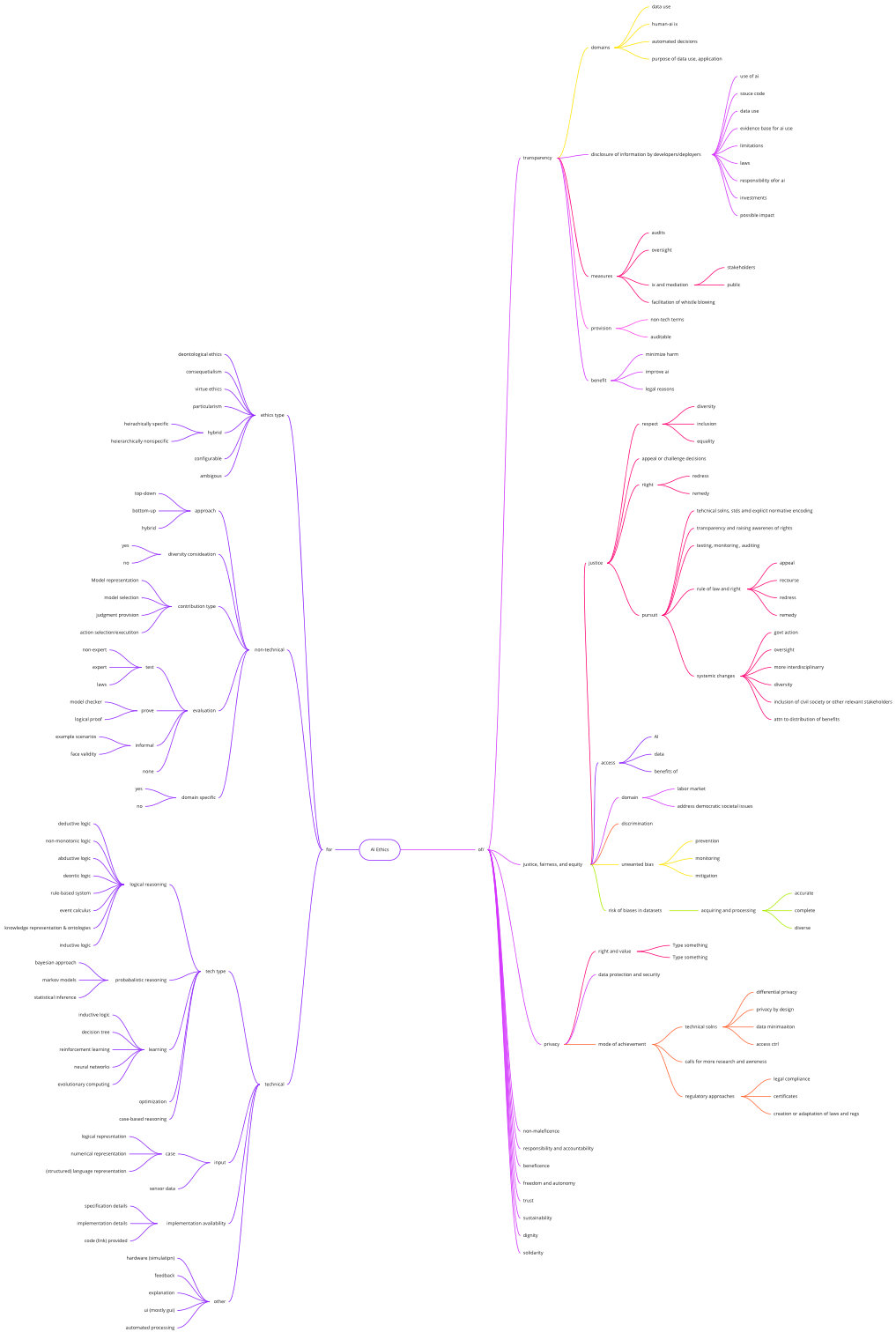

FIGURE 6
Taxonomies, of/as objects vs. for/as subjects.

Therefore, we argue that providing the people insight into when, how, or why the system will fail, will lead to higher levels of trust in autonomy even if (or especially when) the system is less than 99.9% reliable.

Calibration of trust: The discourse on the *calibration* of trust in autonomy most commonly arises from Lee and See's [21] examination of trust in automation. On the face of it, the concept is straightforward: trust in the system should match the system's trustworthiness. However, as we have reviewed, automation does not scale to autonomy, and neither trust nor trustworthiness are unitary—what aspect the trustor's trust is based upon need not match the aspect from where the trustee's trustworthiness is derived. We recommend escaping this complexity by simply replacing trust calibration with reliance calibration: In this way, the axes of calibration would simply be trustor's perceived reliability against trustee's demonstrated reliability. While aspects of the relationship between people and machines are not captured in reliance calibration, the interaction of the person and machine is not aggrandized beyond what the current state of science and engineering can speak to.

We see this aggrandization of *reliability* to *trust* occur, for example, with the reception of findings on algorithm aversion[5]. These findings are typically summarized to claim that people do not expect machines to make errors whatsoever, and so people's "trust" in people is often overrated, whereas people's "trust" in machines is underrated. However, since this phenomenon is almost entirely concerned with performance, it remains squarely within the realm of perceived reliability, and the richness of *trust* may not need to be invoked.

Nonetheless, if the behavior of the system never varies (it performs with perfect reliability), trust is almost irrelevant to the relationship between the person and the machine. For all these reasons, in some cases, the less loaded term of *assurance* (which is licensure-oriented) is more appropriate than the term *trust* (which is state-oriented). For automation, in which action is paramount, and mimicry and rule-following is sufficient (but brittle), the assurance case is based on performance. For machine autonomous systems, in which internal state reflecting the machine's conception of its environment is paramount, and generalization and transfer learning around that environment is possible, the assurance case is based on transparent and interpretable (legible) reasons for why some action was taken over another.

Operationalization and Measurement of trust: Trust is notoriously difficult to measure, in both interpersonal and PMT contexts. As [41] state "a lack of clearly defined measures as they connect to trust theory has also forced scientists to create their own *ad hoc* measures that capture trust as a monolith, rather than a targeted aspect of trust theory." In part, this is due to the phenomenon being a mental state and social relationship to which direct access or quantification is unavailable. Research instead measures proxies from classes including behaviors, subjective assessments, and physiology. However, any of these proxy measures, or even all of them together as a set still do not fully characterize the relevant mental state. The allure that these proxies are measurable drives the conceptualization of trust to meander to meet the proxies. So then, trust is reduced to adoption (behavior), or affinity (subjective assessment), or oxytocin levels (physiology). If we do not measure the right thing, but still optimize for that proxy, are we really saying anything about trust itself? This way of going about science strains the criterion of falsifiability—in these cases, we are searching for our keys under the lamppost, because that's where the light is.

Initial research on trust (of people or machines) relied on subjective measures (e.g., [42]) or indirect measures of trust reflected in the behavior of the person (see for example [43]). Subjective indicators, such as *the negative attitudes toward robots scale (NARS)*, tend to capture more about the likeability of the machine and its position vis-à-vis the uncanny valley or anthropomorphism (eye contact, smiling, nodding, social gesture, responsiveness) than about trust proper. Likeability does not necessarily indicate a willingness to be vulnerable to the machine, especially once the person experiences an event where the machine fails at the task. While such etiquette and immediacy behaviors by the machine are useful to promote adoption, these expressions are manufactured, not earnestly produced as they appear in people, and so designed to manipulate people into a positive disposition, which is not a benevolent affair. When machines produce apologies for poor outcomes, they generally cannot state what they are sorry for, nor can they necessarily change their behavior to ensure that outcome does not occur again (an essential aspect of a genuine apology without which the apology is simply a speech act to get one's way, a sociopathic device). Such a speech act improves perception of the machine's trustworthiness, at least after the first failure, though it is not clear whether repeated apologies would maintain the perception of trustworthiness after a second or third identical failure. Here, notions of betrayal and forgiveness come into play—if these related terms from the interpersonal context seem irrelevant with regard to interacting with a machine, use of the term "trust" must be drawn into suspicion for being just as overzealous.

Physiological indicators of trust are not well established in interpersonal contexts, and it is further unclear whether they would even appear in humans (humans used here instead of people, since the physiology of concern is particular to human biology) trusting machines if those machines are not recognized as social actors, or if teleoperation means that the trust relationship is interpersonal between operator and

5 Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err (upenn.edu), Overcoming Algorithm Aversion: The Power of Task-Procedure-Fit | Academy of Management Proceedings (aom.org)

user, and only mediated by machine. Since the behavior of the trustee also has an impact on the wellbeing or goals of the trustor, that is, there is vulnerability in the act of trusting, the psychophysiology associated with that state may be a more worthwhile measurement target.

Behavioral indicators (including *Acceptance, Deference, tolerance, Workload-resistant compliance, behavioral economics measures like investment*) of trust fail to capture alternatives, and if the options are utilize/adopt or not, then the volitional aspect of trust (willingness) is not being measured. Vigilance/neglect and accepting advice or recommendations are also problematic for the same reason. High workload necessitates neglect and acceptance, which saturates measurement, whereas it is only under these kinds of circumstances that one would employ an autonomous system.

On top of these confusions are another, related to Jobin's divergences mentioned in the previous section on ethics—to what issue, domain or actors does trust and trustworthiness pertain? In the 2020 executive order promoting the use of trustworthy AI in the federal government, most of the principles listed are actually referring to "trustworthy use" instead of "the use of trustworthy AI." Of the nine principles, five are incumbent on the governmental agency to ensure that the institution's *use* is trustworthy—in fact the principle of transparency is reversed from its typical use applying to the technology, and here applies to the governmental agency's transparent use of the AI.

Finally, trust is not only personal and calibrated, but highly contextual—one may trust a particular individual for one task in one context but not in another because as people we have learned the characteristics of the context that suggest potential successful performance. Therefore, the ability of the machine to understand the differences in these contexts, and predict its own performance in the context becomes a useful element for the development of trust between people and robots. In other words, the system may succeed at the task in one instance, performing in the way that the person expects but based on reasoning that differs from the person's basis for action. At a second point in time, the machine may take a different action because the aspects of the context on which it focused are different than in the first instance (while the aspects of the context on which the person focused remain the same). When an autonomous system is created to perform a task, it is designed to achieve the person's goal. When the person performs the task without automation, there are rules that underly how the task is performed—such as to act otherwise could lead to injury. These underlying rules may not be relevant to the machine, as it may not be harmed by the environment as easily. The designer must ask, however, whether the machine should still follow this rule so that the behavior of the machine is more easily predicted, understood, or trusted by the person. Given our conceptualization of trust (and following the argument of [44]) the person in a person-machine team must similarly be able to assess the state of the machine—that is, the ability to assess the risk in teamwork and their own vulnerability to the potential for a mistake by the machine.

9 Conclusion

Words matter—in a very Whorfian way, they shape how we engineer our world. The translation of terms from their original interpersonal use to their use in person-machine teaming contexts must be performed deliberately to maintain conceptual and scientific rigor. The reductive mindset of "human-machine teaming" suggests that a human may be treated like automatons with input and output to be compatible and interchangeable with machines, but in a team or otherwise, machines and people are not equivalent.

This reductive mindset further leads to beliefs that development of machine teammates can ignore the fundamental behavior of people, because the person could just be trained to support the machine. Vice our argument here that people will always be part of the system, "in the loop," "on the loop," or dictating or receiving the output of the autonomous system's actions, we find that too often, the aim in developing autonomous systems centers around the desire to engineer people out of the system. However, this approach undercuts the purpose of developing autonomous teammates. People are social, and will engage in social interactions with entities that have even a modicum of perceived independent behavior. Therefore, person-machine teaming is an inherently social activity, and as such, engineering and development of autonomous systems must acknowledge people as social entities, and account for social behavior in developing the system.

To be of the greatest utility, autonomous machines must be allowed to operate with the initiative and independence they were built to exert. Seeking to control every possible outcome of their behavior reduces them to tools and undermines their usefulness. We must admit that machine performance, just as the performance of people, will rarely be perfect. To that end, in the development of autonomous teammates, we must accept this imperfection and the vulnerability that it entails, to people, to the system, and to the task (see Coactive design (acm.org)). We must acknowledge that development and test environments, even when they are of high fidelity and of adequate ecological validity, will never exactly match the deployment environment of the wild. Instead of controlling machine behavior as a means to achieve some aspect of a trust relationship, we argue that we must appreciate how context affects system performance—both the performance of the machine and of the person. Autonomous machines must not be designed to assume that the person they are teaming with is sufficiently involved in the task to be able to take it over at any time (even with notice), but rather, these systems must be designed for safe and graceful failure that accounts for unmitigable vulnerability. The approach here detailed has significant ethical and legal implications for the development of robots that are categorically different and merit

distinct consideration from those commonly discussed in the development of AI.

Author contributions

AG primarily authored the introduction, and the sections on Semantics and the SMP, Personhood, Vulnerability, Autonomy, and Ethics. JM primarily authored the section on teaming and the conclusion, contributed to all the other sections, and is here celebrated for translating AG's inscrutable language to eschew obfuscation and render the material accessible to the intended audience. AG and JM coauthored the section on Trust.

Funding

This production was supported a JHU Discovery Program grant on *Enabling machines to reason over potential harms to humans* (AG), by APL's *Execution Priority on Ethical AI* (AG), and by the Institute for Experiential Robotics, Northeastern University (JM). The development of the semantic mapping pipeline was supported by JHU's Institute for Assured Autonomy (AG).

Acknowledgments

AG declares special thanks to Joshua D. Baker at APL, and Brian Hutler now at Temple University for the years of discussion that helped hone these thoughts.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Greenberg AM. *Deciding machines: Moral-scene assessment for intelligent systems. Human-machine shared contexts*. Elsevier (2020). doi:10.1016/B978-0-12-820543-3.00006-7

2. Mitcham C. The importance of philosophy to engineering. *Teorema XVII* (1998) 3:27–47.

3. McDermott D. Artificial intelligence meets natural stupidity. *SIGART Bull* (1976) 57:4–9. doi:10.1145/1045339.1045340, no.

4. Turkle S, Taggart W, Kidd CD, Dasté O. Relational artifacts with children and elders: The complexities of cybercompanionship. *Connect Sci* (2006) 18(4):347–61. doi:10.1080/09540090600868912

5. Rieder TN, Hutler B, Debra J, Mathews H. Artificial intelligence in service of human needs: Pragmatic first steps toward an ethics for semi-autonomous agents. *AJOB Neurosci* (2020) 11(2):120–7. doi:10.1080/21507740.2020.1740354

6. Lee JER, Nass CI. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In: *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives*. Pennsylvania, United States: IGI Global (2010). p. 1–15.

7. Nass C, Fogg BJ, Moon Y. Can computers be teammates? *Int J Human-Computer Stud* (1996) 45(6):669–78. doi:10.1006/ijhc.1996.0073

8. Rix J. From tools to teammates: Conceptualizing humans' perception of machines as teammates with a systematic literature review. In: Proceedings of the 55th Hawaii International Conference on System Sciences (2022).

9. Lyons JB, Mahoney S, Wynne KT, Roebke MA (2018). *Viewing machines as teammates: A qualitative study*. Palo Alto, CA: AAAI Spring Symposium Series.

10. Salas E, Dickinson TL, Converse SA, Tannenbaum SI. Toward an understanding of team performance and training. *Teams: Their training and performance*. In R. W. Swezey E. Salas (Eds). Ablex Publishing (1992), 3–29.

11. Walliser JC, de Visser EJ, Wiese E, Shaw TH. Team structure and team building improve human–machine teaming with autonomous agents. *J Cogn Eng Decis Making* (2019) 13(4):258–78. doi:10.1177/1555343419867563

12. Salas E, Cooke NJ, Rosen MA. On teams, teamwork, and team performance: Discoveries and developments. *Hum Factors* (2008) 50(3):540–7. doi:10.1518/001872008x288457

13. Abrams D, Wetherell M, Cochrane S, Hogg MA, Turner JC. Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *Br J Soc Psychol* (1990) 29(2):97–119. doi:10.1111/j.2044-8309.1990.tb00892.x

14. Kozlowski SW, Ilgen DR. Enhancing the effectiveness of work groups and teams. *Psychol Sci Public Interest* (2006) 7(3):77–124. doi:10.1111/j.1529-1006.2006.00030.x

15. Lyons JB, Sycara K, Lewis M, Capiola A. Human–autonomy teaming: Definitions, debates, and directions. *Front Psychol* (2021) 12:589585–15. doi:10.3389/fpsyg.2021.589585

16. McNeese NJ, Demir M, Cooke NJ, Myers C. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Hum Factors* (2018) 60(2):262–73. doi:10.1177/0018720817743223

17. Marble JL, Greenberg AM, Bonny JW, Kain SM, Scott BJ, Hughes IM, Luongo ME. Platforms for assessing relationships: Trust with near ecologically-valid risk, and team interaction. In: *Engineering artificially intelligent systems*. Berlin, Germany: Springer (2021). p. 209–29.

18. Fincannon T, Barnes LE, Murphy RR, Riddle DL. Evidence of the need for social intelligence in rescue robots. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566); 28 September 2004 - 02 October 2004; Sendai, Japan. IEEE (2004). p. 1089–95.

19. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Acad Manag Ann* (2020) 42(2):627–660. doi:10.5465/annals.2018.0057

20. Lyons JB, Sean Mahoney KTW, Roebke MA. *Trust and human-machine teaming: A qualitative study. Artificial intelligence for the internet of everything*. Amsterdam, Netherlands: Elsevier (2019). doi:10.1016/B978-0-12-817636-8.00006-5

21. Lee JD, See KA. Trust in automation: Designing for appropriate reliance. *HFES* (2004) 46(1):50–80. doi:10.1518/hfes.46.1.50.30392

22. Albuquerque N, Guo K, Wilkinson A, Savalli C, Otta E, Mills D. Dogs recognize dog and human emotions. *Biol Lett* (2016) 12:20150883. doi:10.1098/rsbl.2015.0883

23. Tschopp M. Vulnerability of humans and machines - a paradigm shift (scip.ch) (2020). Available at https://www.scip.ch/en/?labs.20220602 (Accessed on August 15, 2022).

24. Roy JL, McAllister DJ, Bies RJ. Trust and distrust : New relationships and realities. *Acad Manage Rev* (1998) 23:438–58. doi:10.5465/amr.1998.926620

25. McDermott P, Dominguez C, Kasdaglis N, Ryan M, Trhan I, Nelson A. *Human-machine teaming systems engineering guide*. Bedford, United States: MITRE CORP BEDFORD MA BEDFORD United States (2018).

26. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev* (1995) 20(3):709–34. doi:10.5465/amr.1995.9508080335

27. Ullman D, Malle B. The effect of perceived involvement on trust in human-robot interaction. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 07-10 March 2016; Christchurch, New Zealand. IEEE (2016). p. 641–2.

28. Ullman D, Malle BF. Human-robot trust: Just a button press away. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction; 6 March 2017; New York, NY, United States (2017). p. 309–10.

29. Berg J, Dickhaut J, McCabe K. Trust, reciprocity, and social history. *Games Econ Behav* (1995) 10123:122–42. doi:10.1006/game.1995.1027

30. Bradshaw JM, Hoffman RR, Johnson M, Woods DD. The seven deadly myths of 'autonomous systems. *Human-Centered Comput* (2013) 2–9.

31. Sheridan TB. Humans and automation: System design and research issues. *Hum Factors* (2002) 39(2):280.

32. Beer JM, Fisk AD, Rogers WA. Toward a framework for levels of robot autonomy in human-robot interaction. *J Hum Robot Interact* (2014) 3(2):74–99. doi:10.5898/JHRI.3.2.Beer

33. Sparrow R. Why machines cannot be moral. *AI Soc* (2021) 36(3):685–93. doi:10.1007/s00146-020-01132-6

34. Müller VC. Ethics of artificial intelligence and robotics. In: EN Zalta, editor. *The stanford encyclopedia of philosophy (summer 2021 edition)*. Palo Alto, CA: The Stanford Encyclopedia of Philosophy (2020).

35. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* (2019) 1:389–99. doi:10.1038/s42256-019-0088-2

36. Asimov I. In: *Run around. I, Robot (The Isaac Asimov Collection)*. New York: Doubleday (1950).

37. Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A. Implementations in machine ethics: A survey. *ACM Comput Surv* (2021) 53:1–38. doi:10.1145/3419633, no. 6.

38. Roff HM, Danks D. "Trust but verify": The difficulty of trusting autonomous weapons systems. *J Mil Ethics* (2018) 17(1):2–20. doi:10.1080/15027570.2018.1481907

39. Atkinson DJ. "Final report : The role of benevolence in trust of the role of benevolence in trust of autonomous systems,(2015). doi:10.13140/RG.2.1.4710.5127

40. Hoff KA, Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum Factors* (2015) 57(3):407–34. doi:10.1177/0018720814547570

41. Kohn SC, De Visser EJ, Wiese E, Lee YC, Shaw TH. Measurement of trust in automation: A narrative review and reference guide. *Front Psychol* (2021) 12:604977. doi:10.3389/fpsyg.2021.604977

42. Schaefer KE. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In: *Robust intelligence and trust in autonomous systems*. Boston, MA: Springer (2016). p. 191–218.

43. Freedy A, De Visser E, Weltman G, Coeyman N. Mixed initiative team performance assessment system (MITPAS) for training and operation. *Interservice/Industry Train Simulation Edu Conf (I/ITSEC)* (2007) 7398:1–10.

44. Hopko SK, Mehta RK. Trust in shared-space collaborative robots: Shedding light on the human brain. *Hum Factors* (2022) 0(0):187208221109039. doi:10.1177/00187208221109039