The Journal of Ethics
https://doi.org/10.1007/s10892-023-09449-2

Automation, Alignment, and the Cooperative Interface

Julian David Jonker[1]

Received: 16 November 2022 / Accepted: 1 August 2023

Abstract
The paper demonstrates that social alignment is distinct from value alignment as it is currently understood in the AI safety literature, and argues that social alignment is an important research agenda. Work provides an important example for the argument, since work is a cooperative endeavor, and it is part of the larger manifold of social cooperation. These cooperative aspects of work are individually and socially valuable, and so they must be given a central place when evaluating the impact of AI upon work. Workplace technologies are not simply instruments for achieving productive goals, but ways of mediating interpersonal relations. They are aspects of a cooperative interface i.e. the infrastructure by which we engage cooperative behavior with others. The concept of the cooperative interface suggests two conjectures to foreground in the social alignment agenda, motivated by the experience of algorithmic trading and social robotics: that AI impacts cooperation through its effects on social networks, and through its effects on social norms.

Keywords AI safety · AI ethics · Social alignment · Value alignment · Cooperative interface · Future of work

Research in AI safety has largely been framed in terms of the problem of *value alignment* i.e. how to align automated processes with human values so that they produce the outcomes we actually want (e.g. Christian 2020; Kearns and Roth 2020; Russell 2019; Bostrom and Yudkowsky 2018; Soares 2016). I will contrast value alignment with *social alignment*, which concerns the impact of automation upon cooperative social networks, social norms, and other aspects of the *cooperative interface*. I will

✉ Julian David Jonker
jonker@wharton.upenn.edu

[1] Legal Studies and Business Ethics, The Wharton School of the University of Pennsylvania, 669 John M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA

Published online: 13 October 2023

clarify and illustrate these concepts, and show that they deserve a special place on the research agenda of AI ethics.

The argument of the paper isn't entirely conceptual, since the content of *value alignment* is not straightforwardly produced by our conceptual capacities as revealed by linguistic intuitions, but instead emerges from a research practice i.e. the particular research questions, approaches, and theories that articulate a distinctive problem with that name. Therefore one premise of the argument is practical: research in AI safety as it is currently practiced has left out an important concern, namely *social alignment*. Social alignment takes seriously the impact that even well-aligned machines might have upon social cooperation, since it considers the social and institutional responses to AI to be a key part of AI's cooperative impact. There is also a conceptual premise: social alignment is not simply value alignment with the assumption that cooperation is our goal. Finally, the argument has a normative premise: social cooperation is an important goal. The conclusion of the argument is that social alignment research should be given a prominent place on the AI safety research agenda, since it raises questions that are either unlikely to be considered or cannot be considered by the value alignment agenda.

This conclusion is a general one about AI ethics, but I will often focus on the impact of AI upon work. This is a particularly relevant domain for focusing the argument, given that work is an important but sometimes neglected site of social cooperation, as I describe in Sect. 1. Section 2 turns to the conceptual and practical distinctiveness of social alignment. This distinctiveness is best understood by consideration of what I will call "the cooperative interface" i.e. the complex of norms, practices, and technologies that allow us to engage in cooperation with each other. Given that affordances for interpersonal cooperation emerge from this complex, the alignment of a new technology with our preferences does not entail that the cooperative interface will be positively impacted. Section 3 generalizes the point that automation and AI in the workplace may negatively impact social alignment. It presents two conjectures: that automation restructures social networks in a way that changes the distribution of cooperation, and that it erodes norms of interpersonal cooperation. Section 4 concludes by showing that social alignment is practically and conceptually distinct from value alignment, since it raises questions about the cooperative interface that are not simply questions about the alignment of particular technologies with human values.

1 The Workplace as a Cooperative Institution

In this section I argue that AI safety should be concerned with its impact on cooperation, including (as an important case) its impact on cooperation at work.

I will primarily be concerned with showing that cooperation at work is a good means for establishing *basic social cooperation* i.e. the kind of cooperation that is the basis of a stably fair society. As such, cooperation at work is merely instrumentally valuable, but this does not undermine its value, given the intrinsic importance of basic social cooperation. Until we have guaranteed better ways to establish basic social cooperation, we should be mindful of how technological change affects our current ways of establishing it.

Why think that basic social cooperation is an intrinsic good? Claims about intrinsic value are hard to defend and there is no room here to defend such a foundational assumption. But the assumption can be motivated by pointing out that the value of basic social cooperation is implicit in the social contract tradition that is the central tradition of modern political philosophy. Hobbesian contractarianism, one of the central forms of the social contract tradition, is premised on the idea that we have reason to accept the authority of morality and the state because we all stand to benefit from cooperation with each other (e.g. Narveson 2001: 142–159). And in the contractualism of John Rawls, which played a defining role in the resurgence of political philosophy in the late twentieth century, the very idea of a just society is that of a fair system of cooperation (Rawls 1999: xv, 88). Following this Rawlsian idea, I take basic social cooperation to be the condition of possibility of any social ideal worth pursuing. This does not provide an argument that we should maximize or even increase present levels of cooperation, but it does suggest that we should be wary of threats to present levels of cooperation. Such wariness is a sufficient normative starting point for this paper's argument.

Why should we care about cooperation *at work*? The workplace is a cooperative institution in the sense that it is part of the larger cooperative infrastructure of a democratic society and advances basic social cooperation. Workplace cooperation is at least instrumentally important because of its connection to basic social cooperation. The importance of basic social cooperation is an assumption that runs through the mainstream of political philosophy: such cooperation is the basis for the other goods of society.

There are at least two important dimensions of workplace cooperation, the productive and pedagogical. Though they are related in practice, I distinguish these dimensions of cooperation since it is the pedagogical dimension of workplace cooperation that has special ethical significance through its connection to basic social cooperation.

The *productive dimension* of workplace cooperation refers to the way in which cooperation at work is a crucial aspect of economic production. Indeed, the reason there are workplaces in the first place is the same reason that there are firms: to enable the sort of cooperation that facilitates economic production. As noted by the theory of the firm, there are a variety of transaction costs that would be an obstacle to profitable forms of production taking place by way of market allocation, but that can be overcome by organizing production within the firm (Williamson 1973, 2010; Klein 1988; Rajan and Zingales 2001). Firms arise as an equilibrium outcome of the efforts of individuals to reach optimal contracts with each other against the background of agency costs, including the costs to a principal of monitoring the agent's behavior (Jensen and Meckling 1976; Alchian and Demsetz 1972). The collocated workplace, and the productive aspect of cooperation associated with it, can also be viewed as an equilibrium outcome of attempts to overcome such agency costs as monitoring and shirking (Yellen 1984).

It is tempting but misleading to think that the kind of cooperation suggested by this economic theory of the firm is so thin that it does not deserve the term. Indeed, philosophers who study work commonly emphasize the role of hierarchical authority in the workplace, and the moral deficiencies of such hierarchy (Hsieh 2005; Anderson 2017; Frega et al. 2019; Tsuruda 2020). The philosophical emphasis on authority echoes the original conception of the firm in economic theory (Coase 1937), which assumed that insofar as employees are subject to the authority of managers who are superior to them in a chain of command, these managers will have the power to monitor workers and to direct them in minute detail, thus overcoming transactions costs that would be an obstacle to competitive production. Yet the economic theory of the firm has evolved beyond this emphasis on hierarchical authority. The influential team production model of the firm (Alchian and Demsetz 1972) points out that the productive efficacy of the firm is due not solely to hierarchical authority, but to the fact that the firm involves a network of contractual relations in which a central node of the network is able to coordinate production as a team.

The *pedagogical* dimension of workplace cooperation refers to the way in which working together teaches us the more general skill of civic cooperation i.e. living together as members of a political community. There are at least three ways in which we learn valuable skills at the workplace that directly transfer to our ability to live well together as a political community.

First, the workplace requires that we work peacefully and productively, and often closely, with strangers with whom we are likely to disagree, and who likely have different demographic features and social positions, and divergent interests. In this way we learn the skills of working together with strangers and accommodating divergent viewpoints and behaviors. Now, the degree to which our workplaces tolerate difference and encourage diversity is no doubt disappointing to many of us. Yet it is still notable, and no accident, that civil rights legislation in the US has taken the workplace as a site of special significance. For example, Title VII of the Civil Rights Act of 1964, which targets discrimination and harassment in the workplace, has become a model for other antidiscrimination legislation, as has the Americans with Disabilities Act of 1990, which also places great emphasis on workplace accommodation of people of all abilities. These legislative initiatives wisely use the workplace as a lever for achieving social integration, because most people must work, and most workers must work with strangers. Therefore workplaces are where people are mostly likely to encounter difference, and are good places for the state to encourage the virtues of diversity and toleration (Estlund 2003).

Second, the workplace establishes valuable social networks that further our cooperative capacities. The sociologist Mark Granovetter's (1973) work on weak ties showed that networks made up of relatively low-intensity social relations can be important in the diffusion of ideas and knowledge. Burt (1992, 2005, 2010) has furthered this project of social network analysis by showing how an individual's position in a social network constitutes valuable social capital. Many of his examples are drawn from the world of work, unsurprisingly so given how important social networks are to our working lives. It is also well documented that social networks are an important way that people hear about job opportunities and obtain jobs (e.g. Granovetter 1995). Yet the reach of a job seeker's social network may depend upon their race and gender (Pedulla and Pager 2019; MacDonald 2009). Therefore one reason to integrate workplaces is to foster diverse social networks that will enhance economic opportunities for all (Anderson 2010).

Third, some forms of work enhance cognitive and conative capacities (such as the ability to take in complex forms of information, engage in rational deliberation with others, and exercise autonomous choice in a responsible manner) that are in turn valuable for exercising the civic capacities that enable a well functioning democracy (Satz 2023). In their early study of comparative political culture, Almond and Verba argued that the experience of authority and autonomy at work has an influence upon workers' capacities to engage in the political system, such as willingness to make appeals to government officials or form lobbying groups (1963: 363–366). Satz gathers similar evidence suggesting that the complexity and hierarchical nature of work can have an effect on workers' cognitive abilities, self-esteem, emotional intelligence, and self-direction outside of work, concluding that "[w]ork, whatever else it is, seems to be a giant school" (Satz 2023: 21).

In sum, the workplace is a cooperative institution both in the sense that its productive efficiency is based on cooperation between workers, and in the sense that it is part of the larger cooperative infrastructure of a democratic society. Its cooperative nature is a valuable part of our social arrangements as they presently are. To this the objection may be raised that cooperation in the workplace is not an essential ingredient of social cooperation, since if we were to do away with work other institutions that better enable cooperation might take its place. I am happy to concede this possibility. My argument is not that work is metaphysically required for social cooperation, but simply that in our present form of society it plays an important role in educating workers in the forms of social cooperation that we value. In this respect it is the pedagogical dimension of work that is particularly important, as we can see if we contemplate how we would replicate the function of antidiscrimination legislation in a world without work. Perhaps we could replace this function with due thought and effort, but we should not expect it to survive technological change automatically. This sort of threat is enough to give normative grounds for the research agenda I propose here.

2 The Cooperative Interface

A focal point for thinking about the impact of automation upon social cooperation is what we might call the "cooperative interface". In what follows I'll give several examples of the cooperative interface in different contexts: traffic intersections, office doors, and assembly lines. More abstractly, the cooperative interface is the infrastructure that determines the opportunities for cooperation, the affordances that allow individuals to engage in cooperation, and the styles of cooperation that are salient. The importance of this idea is its abstraction. Cooperative styles and affordances emerge from the complex interaction of social norms, organizational forms, and technological and material architectures. The idea of the cooperative interface allows us to move between thinking about the broad impact of technological change upon basic social cooperation and the specificities of how technology shapes cooperation in highly specific contexts, such as particular workplaces. In particular, it connects empirical questions about technological change with normative and conceptual questions about the kinds of cooperation we value. The section will end by highlighting how the idea of the cooperative interface is continuous with the discipline of human-computer interaction, while also going beyond it.

We are familiar with the concept of an interface from computing: the command line is one way of interacting with a computer, the point-and-click graphical user interface is another. In fact, all of our tools have interfaces i.e. ways of interacting with them in order to activate their functions. The idea of a *cooperative* interface rests on two further thoughts. First, social cooperation involves forms of interaction, i.e. there are particular ways in which we activate cooperation and interact with others' cooperative behaviors. Second, our machines (as well as our organizational structures and social norms) are elements of these forms of interaction and thereby shape the possibilities for activating social cooperation and determine which of these possibilities appear ready to hand. Accordingly, machine interfaces can be more or less cooperative in the sense that they make it relatively easier or harder to engage in cooperative behavior with others, and they make available different styles and parameters of cooperation by presenting different affordances for engaging cooperative behavior with others. In this way they help to determine the interface by which we cooperate with others.

Let me illustrate these ideas in a familiar setting. Imagine two intersecting roads, and drivers proceeding in different directions along each road. One driver wants to continue in her current direction, another approaches from the opposite direction and wants to use the intersection to turn across the first driver's path. Each wants to proceed as quickly as possible, but none wants to collide. How can they cooperate in doing so? First, consider the design of each driver's car. Likely they each have turn signals i.e. lights that display that a driver intends to turn left or right. The fact that one driver can show his intention to turn makes it easier for the other driver to cooperate by slowing down and yielding. The stalk or lever that the driver uses to turn on these lights is part of the *car's* interface—it is his way of interacting with the system of the car to activate the turn signal. But the light itself, which is really to say the signaling system of which it is the manipulable part, is an aspect of the *cooperative* interface—it is the way in which the turning driver engages in a cooperative interaction with the approaching driver. A driver who lacks such a signaling system would have to find another way of engaging cooperation, such as using hand signals or reading intent from facial and behavioral cues. In that case, the fact that each has a clear windscreen rather than a tinted windscreen would make cooperation more or less readily available—suggesting that the transparency or opaqueness of a vehicle should be regarded as another aspect of the cooperative interface.

It is not just the engineering of our cars that makes cooperation on the road possible. The design of the roads plays an important role too. A variety of designs is available to allow our drivers to coordinate their behavior where their paths cross: an intersection with traffic lights, a roundabout, a four way stop. Each of these will have different effects upon cooperation because of the opportunities for interpersonal interaction they make available and salient. The four way stop, for example, will require more of each driver's attention than the intersection with traffic lights, but it may be more efficient at times of low traffic and therefore reduce frustration. And note that the idea of a cooperative interface does not single out the design of the roads or of the car, but treats these as working together. The combination of turn signals with a roundabout constitutes a different interface than the combination of turn signals with a four way stop. Therefore institutional design—road planning and traffic law, in this case—is as important a part of the cooperative interface as the engineering of the machines we use on the road, and it is particularly important to think about how these elements work together.

There is a more or less empirical question about the optimality of cooperative interfaces—whether some car-and-road architectures interfaces will make it easier to activate cooperative behavior than others. But there are also questions about the design of a cooperative interface that are not entirely empirical. One such question concerns which style of cooperation we prefer: do we want the cooperative interface to emphasize the autonomy of drivers, so that much depends on their choice whether to cooperate or be polite (such as at a four way stop)? Or do we prefer to foreground the safety of drivers over their autonomy (as in the case of traffic lights)? Do we care more about cooperation between drivers who interact at the intersection than we care about their feelings toward the generalized population, such as the feeling of frustration one might feel at a slow traffic light? These are normative questions about the style of cooperation we value in a context. There are also questions about the pedagogical effects of the cooperative interface, i.e. the impact that it has on the cooperative dispositions of drivers and others. Do we want drivers to frequently engage in effortful cooperation at intersections in the hopes that they will be habituated into being more cooperative drivers elsewhere? This is partly an empirical question about how a driver's character is likely to be shaped by a particular interface, and partly a normative question about when and why we care about cooperation on the roads, and whether we care about individuals having cooperative virtues.

These are questions of political philosophy in part because they have a normative element, but also because they are about the general structure of society, to the extent that the cooperation local to the cooperative interface (say: cooperation on the roads) is connected with cooperation in the rest of society. The fact that drivers are less prone to road rage may well lead to less fear of strangers in society at large, and a more polite culture on the road may both reflect and instill a more polite culture in the town hall. Whether this is in fact so is again an empirical question. But if there are such connections, then there are philosophical questions about how to value increased cooperation in one area of life given its effects upon cooperation elsewhere and given the tradeoffs in other values such as efficiency and autonomy.

The example of the cooperative interface of driving is an illustration of a general point: our machines and institutions can be more or less helpful in enabling us to engage in cooperative interactions with each other, and they can affect our cooperative style and dispositions. For an example closer to our topic of work, consider the humble office door. The office door, when shut, is helpful to an office occupant who wants privacy or undisturbed focus time. When open, it is helpful to the occupant as well as her colleagues because it facilitates communication. Indeed, it enables communication rather than simply allowing it: the fact that the door could easily have been shut makes an open door into a signal that interaction is welcomed. The office door is in this way part of the cooperative interface of the workplace. It presents an affordance for engaging in a particular style of cooperation, one that proceeds by way of signaling communicative and collaborative intentions, and so it is not simply a neutral instrument in the cooperative landscape.

By contrast, the open office encourages more frequent encounters with colleagues, but by disabling the signaling function of the office door it also facilitates a different cooperative style, one that may result in more extreme non-cooperative behaviors such as avoiding the office altogether or working with headphones on. Studies suggest that open offices lead (against intuitive expectations) to a decrease in face-to-face interaction in favor of electronic communication (Bernstein and Turban 2018). This fact does not mean that there is *less* cooperation in an open office. That will depend on what behaviors count in a particular context as cooperative. Thus exploring the cooperative interface is not a purely empirical task. Different artifactual and institutional design present different affordances for cooperative behavior, encouraging individuals to engage (or not) in cooperation according to different styles. Studying the cooperative interface is partly conceptual and normative insofar as we must articulate and evaluate the cooperative styles made available by different interfaces.

The same point can be made by considering the assembly line as an essential part of the cooperative interface of the industrial factory. On the one hand, the assembly line makes possible a massive increase in cooperation of a relatively thin sort. By giving factory workers specialized roles and using automation to organize these roles into a cohesive process, it is possible to coordinate many more workers than previously possible and thereby achieve large economies of scale (Ford 1922). On the other hand, the assembly line isolates workers, makes it harder for them to interact with each other socially beyond the coordination scheme imposed by the assembly line, and makes their work feel less meaningful in the broader scheme of their social lives (Walker and Guest 1952). It is not trivial to say that the assembly line has increased or decreased the amount of cooperation in the factory. All that is evident is that there has been a qualitative change whose evaluation requires further conceptual and normative inquiry into the forms of cooperation we value and the tradeoffs we are willing to make.

These examples—the office door and the assembly line—should not be understood as purely technological or physical phenomena. They combine the material aspects of the workplace with social and organizational aspects of work. So the cooperative interface of the workplace should be seen as an abstraction that emerges from the interactions of physical architecture (e.g. whether the workplace is an open office, an office with doors, a remote office, or some other design); organizational structure and social and legal norms (e.g. whether the workplace is unionized, flat or hierarchical, and what sort of informal social culture it has); and the kinds of technology used in the workplace (e.g. whether workplace participants work with dangerous machines, whether they communicate in person or through synchronous or asynchronous messaging).

Given this complexity, it is difficult to predict a priori what the impact of a particular technology will be in a particular organizational context, and we should not expect the relation between AI and cooperation to be straightforward. Certainly, the introduction of computer-mediated communication has not had a straightforward impact upon cooperation between humans. Consider a study by Kiesler et al. (1996) that showed that people were more likely to cooperate with other human beings in a Prisoner's Dilemma setting than with computer partners. The study nonetheless found that cooperation with computer partners increased where the computer communication was more human-like i.e. used a human voice or synthesized human face. This suggests that there are parameters that affect cooperative behavior when it is mediated by computers, and indeed subsequent research has found that different modalities of computer-mediated communication do differentially affect cooperation (Jensen et al. 2000; Brosig et al. 2003). These modalities appear to affect the quality of communication, its synchronicity, and therefore the salience of social norms of cooperation (Bicchieri and Lev-On 2007). In addition, recent work in consumer psychology shows that people will disclose more about themselves on their smart phones than on their laptops (Melumad and Meyer 2020). This sort of finding suggests that there are more relevant factors than simply the modality of communication, but also (for example) the attachment an individual has built with certain pieces of technology (Melumad and Pham 2020; Fullwood et al. 2017).

In light of such non-linearities, the idea of the cooperative interface provide us with a framing device for inquiries into the cooperative impact of different combinations of technological circumstances and institutional design. The idea of the cooperative interface is particularly appropriate for thinking about the impact of AI, given that the idea of an interface has already been so important in the history of computing in general, and the field of human-computer interaction (HCI) in particular (Erickson and McDonald 2007). HCI emphasizes the concept of *interaction*, previously reserved for interpersonal relations, and applies it to human-machine relations (Suchman 2012: 34). The project of artificial intelligence has itself been put in the language of HCI by being described as a way of enabling "mutual intelligibility" of humans and computers (Suchman 2012: 31).

The research program I aim to articulate here extends this focus on interaction: but it is not primarily about mutual intelligibility of human and machine, but rather how machines mediate human cooperation. The program fits naturally with contemporary insights in the philosophy of action. Human action is distinctive insofar as it is guided by intentions, which we can think of as the elements of planning (Bratman 1987). The planning conception of intentional action can be extended to the shared activities we undertake together with other human beings, insofar as cooperative action involves plans that refer to the intentions of others (Bratman 1992, 2014). But our plans can also refer to the states and operations of machines and other artifacts, and these artifacts can also occlude or present the intentions of others for cooperative activity. The cooperative interface highlights this nexus of interpersonal attitudes and machine mediation. It therefore shifts our attention beyond the question of mutual intelligibility with machines and toward the question of how machines mediate the mutual intelligibility of one human being and another.

3 Automation and the Cooperative Interface

The idea of the cooperative interface focuses our attention on the impact of AI and automation on interpersonal cooperation. The cooperative functions of the workplace outlined in Sect. 1 moves us to ask about this impact on the workplace in particular, and with special urgency. This section will explore some ways in which AI and automation might impact cooperation at work. It presents two conjectures about the cooperative impact of AI and automation upon social networks and social norms, as examples of the kind of inquiry brought into focus by the cooperative interface. Each conjecture is motivated by an example of a negative cooperative impact of AI: increased secrecy and opacity in the case of the automation of financial markets, and the entrenchment of gender stereotypes with regard to domestic and servant roles by the anthropomorphizing of AI agents. But there is no reason to limit our attention to such negative impacts only, especially as we aim to improve our existing cooperative interfaces.

Note that I will talk of AI *and* automation so as to avoid the risk of conflating these (as talk of *robots* often does). I follow economists Agrawal et al. (2018) in thinking of AI as a class of techniques for making predictive judgments, and automation as the replacement of humans in a given workflow. There is a close connection insofar as AI often leads to automation of some aspect of workflow. But human replacement is not a necessary outcome, since the increased ability to make effective predictions by using AI may also create opportunities for human beings to make these predictions useful by adding their capacity for non-predictive judgment, decision-making, and action. Analogously, the development of spreadsheets made arithmetical calculations cheap and quick, but did not do away with the need for bookkeepers. Instead it created opportunities for the kind of analysis and judgment that spreadsheets could not do on their own, but that humans equipped with spreadsheets could do (Agrawal et al. 2018: 141–142). Still, there are cases in which AI is likely to change a workflow in ways that make it easier to replace human beings—according to some estimates, as much as half of current employment is threatened (Frey and Osborne 2017; cf. Eloundou et al. 2023; Susskind 2020: 91).

Nonetheless, the central concern for our attention here is not that AI will erase some jobs. Rather, it is the impact that AI and automation might have on human-human cooperation, and the fact that design principles for tailoring this impact have not yet been the focus of specialized disciplinary attention. Indeed, we have already been blindsided by the impact that AI has had on cooperation in the political sphere, given that AI plays a central role in the governance of the social media platforms that have played an important part in the political polarization of the past decade (Tufekci 2018; Lorenz-Spreen et al. 2022; cf. Brown et al. 2022, Chen et al. 2022). Though it does not foreground AI, Zeynep Tufekci's (2017) study of the use of social media by 21st century protest movements is an exemplary examination of how we might think about the automated cooperative interface. Tufekci describes the distinctive affordances of social media, showing how they at once enable swift decentralized organization and fail to foster the depth of cooperation that accompanied older modes of organizing. We should consider the impact of AI and automation with a similar attention to its cooperative affordances, examining both the novel opportunities for cooperation and the distinctive style and limits of cooperation that these new technologies foster.

As an illustration of how we might pursue this research agenda, I make two conjectures about the sorts of cooperative impact we can expect AI to have. These are both empirical claims, and while I will motivate them here I do not purport to present empirical confirmation of them, given that the aims of the paper are conceptual and normative. In addition, the conjectures do not make claims that any impact is essential to AI. Rather, we should expect the social impact of AI to be the result of a social process that is determined in part by the nature of that technology and in part by the existing social environment.

The first conjecture concerns AI's *impact on social capital.* The second concerns AI's *impact on social norms*. Social capital, understood as the advantages gained from one's position in a social network, is an important determinant of whether, how, and with whom human beings will cooperate (Burt 1992, 2005, 2010). Social norms, understood as the generally shared expectation that people will behave in a certain way and deserve criticism for not acting in that way, is one of the enabling circumstances of cooperation (Bicchieri 2006). Therefore understanding AI's impact on these dimensions of cooperation is an important way of understanding how AI shapes our cooperative interface.

My first conjecture is that AI and automation radically redistribute social capital by reintermediating social networks, typically reducing the need for thick interpersonal relationships within an industry or market and therefore reducing the social capital of those who would have brokered these relationship. This amounts to a reduction of cooperation of the kind we see in thick interpersonal relationships, though it may increase other styles of cooperation by making market or institutional opportunities more widely available.

An example of this sort of redistribution of social capital by reintermediation could be seen in the financial services industry, starting with the introduction of electronic trading in the 1980s and culminating in the algorithmic high-frequency trading (HFT) that is a prominent part of markets in the early 21st century (MacKenzie 2015, 2017, 2018, 2021; Lange 2016; Zaloom 2006). Before the 1980s, trading was a very embodied and social activity. As reported in Donald Mackenzie's ethnographic work on the futures open outcry pits of the Chicago Mercantile Exchange (CME), trading rewarded those who were tall and those who could yell, i.e. those who could make eye contact with brokers and enter their trades before others (Mackenzie 2021: 44–47; Melamed 2009: 26). This physicality was accompanied by thick interpersonal relationships. in which reciprocity and reputation mattered (Mackenzie 2021: 45).

The physical and social architecture of the open outcry pit was so taken for granted that the first proposal for electronic trading systems attempted to replicate the look of a trading pit, identifying orders with named avatars who could not be present in more than one electronic pit at a time (MacKenzie 2021: 50). Today the idea of trading by clicking on someone's named avatar in an electronic representation of a trading pit will seem absurd to anyone who has traded on an electronic exchange with a central order book. The social assumptions of trading, and the opportunities available to participants, have changed dramatically. Orders can be placed from anywhere in the world with the click of a button and without knowledge of who one's counterparties are. Given this impersonality, one may as well have a bot as one's counterparty—and very frequently this is what the ordinary retail investor or trader has for a counterparty, given that around half of trading volume on US exchanges is due to high-frequency trading (HFT) conducted by algorithms (Zaharudin, Young, and Hsu 2021).

HFT is the culmination of the digitization and automation of financial markets. Algorithms compete, interact, and learn from each other in dark pools (exchanges open only to large institutional clients with opaque order books) with unknowable outcomes in a sort of electronic Darwinian ferment (MacKenzie 2019). HFT shops employ a variety of strategies for intraday trading, but many of them rely on speed and opaqueness to arbitrage differences in speed and information that arise from the flow of trades (Aquilina et al. 2022; MacKenzie 2021; Lewis 2014; Patterson 2012). But behind the impersonal workings of algorithms is a social practice that emphasizes conflict over cooperation, disembodied engagement with anonymous opponents, and ostentatious displays of secrecy even within trading firms (Lange 2016; Preda 2013; Knorr Cetina 2009; Knorr Cetina and Bruegger 2002).

In comparison, older styles of trading that have continued in other markets retain their relatively thick and cooperative forms of sociality. For example, in markets for treasuries in the US and Europe, participants assume fixed roles. Banks that wish to buy treasuries must do so through dealers, who may trade with each other only through interdealer brokers. This dealer-client structure contrasts with the all-to-all structure of the futures and equities markets that have been more thoroughly automated (MacKenzie 2021: 107–114). In all-to-all markets, trading is intermediated by an anonymous order book that can be accessed by all participants regardless of role. The result is not a complete lack of cooperation, but rather a distinctive cooperative style. An anonymous and freely accessible order book allows anyone to transact in the marketplace; but it is also the enabling condition for the trading algorithms which compete against each other in speed and opacity and end up imposing costs on retail investors (Aquilina, Budish, and O’Neil 2022).

Financial trading is not the only domain to exhibit this shift in cooperative style. Consider the transformation of the advertising industry under the influence of the internet, which has largely been a matter of the automated and algorithmic buying and targeting of adverts online. Tim Hwang has compared the resulting changes in the social architecture of the advertising industry to those seen in finance, remarking upon the similar ways in which they have evolved from long term face-to-face relationships between clients and advertisers to a open-ended and impersonal market structure (Hwang 2020: 53–54). In both cases the new structure is not exactly disintermediated or decentralized, but rather involves the replacement of long term personal relationships with intermediaries that have an arm’s length (i.e. impersonal) relationship and a larger appetite for risk (Rajan 2006). Neither Hwang nor Rajan draw a direct connection between AI and the processes of disintermediation and re-intermediation they describe, let alone say enough to underwrite a causal connection. But the striking correlation is enough to motivate investigating my first conjecture.

My second conjecture is that the introduction of AI changes the social norms governing interpersonal interactions so that our interactions with each other are more consistent with our interactions with AI agents. The motivation for this conjecture is twofold: first, there is mounting evidence that we anthropomorphize robots and other AI agents (Marchesi et al. 2019; Złotowski et al. 2015; Bartneck et al. 2009); second, there is mounting evidence that our interactions with other human beings are driven by implicit biases instilled by our social environment (Huebner 2016; Dasgupta 2013; Greenwald et al. 1998; Banaji and Hardin 1996). The conjecture assumes that when we anthropomorphize AI agents we change our social environment, especially as these agents become more pervasive, and the changes in our social environment impact the implicit biases we bring to human interactions.

As an example, consider the entrenchment of gender stereotypes in our interaction with robots. We know that people are prone to gendering robots on the basis of their external features, such as whether they have grey or pink lips (Powers et al. 2005), or long or short hair (Eyssel and Hegel 2012); voice (Crowell et al. 2009; Powers et al. 2005); as well as their function and whether it is a gender-stereotypical tasks such as security or health care (Tay et al. 2014). These gender identifications lead people to make gender-stereotypical judgments about the emotional intelligence of the robot (Chita-Tegmark et al. 2019), and differential judgments about willingness to interact (Kuchenbrandt et al. 2014), and the robots' credibility, trustworthiness, and engagement (Siegel et al. 2009). The most prominent AI agents on the market before the release of large language models like ChatGPT were Amazon's Alexa and Apple's Siri, and these widely available agents were marked through naming and voice as female-gendered. Given that implicit biases are learned from our social environment, these facts raise concerns that the way we gender our robots and AI agents will feed back into the way we gender other humans, and in particular that it may impact our gender-based expectations of and dispositions toward other human beings.

My aim here is not to go beyond these circumstantial facts and give evidence for the empirical claim that AI impacts social norms concerning gender-appropriate tasks. But the circumstantial facts are enough to motivate a research agenda that focuses on AI's impact on social norms. Note that the focus of this agenda would be holistic. The question it poses is not about the engineering of particular AI agents or automated processes, but rather about the way in which these agents and processes are embedded in institutional and social contexts. The decision to give an AI agent a gendered name is not a matter of the engineering of its algorithm, but a matter of how it is embedded in the social world. As I argue next, it is this last point that distinguishes the agenda I am presenting here from the one that has so far defined AI safety.

4 Putting Social Alignment on the AI Safety Agenda

I have presented two conjectures about the cooperative impact of AI, without attempting to confirm them empirically. That is because my aim is instead conceptual and normative. I wish to motivate a research agenda which treats the introduction of AI and automation into the workplace as a social process, and considers the question of AI safety as a broad normative question about how that social process should go, rather than a narrowly technical question about how to engineer AI agents so that they do what we want them to. In what follows I will distinguish that agenda from value alignment, the most prominent line of AI safety research today.

I'll call the agenda I wish to defend the "social alignment agenda." This should be distinguished from what Gabriel and Ghazavi (2021) call "social value alignment," which is simply a call for the techniques of value alignment to take into account the values of all, rather than single users or local technological elites. Social alignment, by contrast, aims to ensure that the introduction of technology into the social environment has a desirable impact on the cooperative interface.

The conjectures I have motivated are likely to occupy a prominent position on this agenda, though it will require further empirical and conceptual work to diagnose the precise ways in which automation and AI impact the cooperative interface, and it will require further normative work in light of empirical findings about this impact to determine the appropriate methods for engineering the cooperative interface that we want. In this final section I want to show that the social alignment agenda is an important addition to the research agenda of AI ethics and AI safety, which has largely been framed as a problem of *value alignment*, even when the term itself is not used (Taylor et al. 2020; Christian 2020; Kearns and Roth 2020; Russell 2019; Bostrom and Yudkowsky 2018; Soares 2016). So I argue now that social alignment is distinct from value alignment, and that the social alignment agenda points to important questions and phenomena that would not naturally be foregrounded by the value alignment agenda.

To begin, I assume that a *research agenda* in this domain comprises a hierarchy of general practical goals, as well as a set of questions, hypotheses, and conjectures whose investigation appears helpful to furthering the practical goals. At any given time, researchers will likely emphasize particular research methods, bodies of disciplinary knowledge, and fields of inquiry in their investigation of the relevant questions, hypotheses, and conjectures. Of course a research agenda is a dynamic entity, since certain avenues of inquiry are likely to appear less fruitful than first expected, and researchers will venture out along new paths. But at any given time it is possible to evaluate a research agenda by considering either whether its practical goals are in fact appealing, or whether its particular questions and methods are a promising way of pursuing the goals.

The chief practical goal of the value alignment agenda is to make AI safe for humans. A more precise subsidiary goal is to identify those AI techniques that will produce outcomes consistent with the values that are important to humans. These goals are explicit in the literature. Value alignment gained prominence within discussions about whether we should fear the future development of human-like artificial *general* intelligence (AGI) that would destroy humanity by accident or as a pre-emptive strike (Bostrom 2014). But more mundane contemporary concerns about AI ethics and AI safety are also increasingly understood along the lines of the value alignment agenda—for example, concerns about algorithmic bias and privacy (Christian 2020), or the risk of harm from poorly designed machine learning systems (Amodei et al. 2016), or the explainability of complex automated systems (Bobu et al. 2023). In each case, the question is whether the preferences and goals embedded in an AI system—for example, in the reward function of a machine learning algorithm—are compatible with the preferences and goals of its users and other stakeholders.

As an illustration, value alignment researchers have pursued the following sorts of questions, methods, and problems (this is not by any means an exhaustive list):

- Tampering/Reward Hacking: can an AI agent be given a proxy reward function that does not result in unintended behavior that leads to a worse outcome relative to the true reward function specified by user intentions? (Hibbard 2012; Kumar et al. 2020; Everitt et al. 2021; Skalse et al. 2022)
- Corrigibility: can an AI agent be guaranteed to be open to human intervention when it performs in a misaligned way? In particular, is an agent guaranteed to be capable of

being shut down? (Soares et al. 2015; Hadfield-Mennell et al 2017; Wängberg et al. 2017; Holtman 2019)

- Inverse Reinforcement Learning (IRL): Specifying our values explicitly is hard (Soares 2016). Can an AI agent instead be aligned by learning human values implicit in data about human preferences i.e. can we use unsupervised learning to discover the reward function that we want to use to align the agent? (Ng and Russell 2000; Hadfield-Mennell et al. 2016; Leike et al. 2018; Arora and Doshi 2021).
- Reinforcement Learning by Human Values (RLHF): RLHF, which has come to prominence because of its use in training ChatGPT, involves fine-tuning an AI model by way of human feedback about the appropriateness of its answers. Is this a useful form of alignment? (Christiano et al. 2017; Bai et al. 2022; Ganguli et al. 2023)
- Power Seeking: it is rational for a wide class of agents to seek to enhance their power whatever their reward functions; what risks are entailed by such power seeking, and can we build AI agents that do not seek power? (Cohen 2020; Carlsmith 2022; Krakovna 2023)

While the idea of value alignment is capacious, it is clear from these paradigmatic examples that the literature focuses on technical interventions into the engineering of various classes of AI agent. By contrast, questions of social alignment focus on the cooperative impact of AI, and are more likely to draw upon the social sciences and political philosophy for their hypotheses, diagnoses, and prescriptions. We have already seen, in the conjectures of Sect. 3, the sorts of issues that are brought into focus by the social alignment agenda. More generally, there are likely to be two areas of particular importance for social alignment:

1. *What is the impact of AI and automation on interpersonal cooperation? How can positive impacts be guaranteed and negative impacts avoided?* The conjectures about social networks and social norms fall within this line of inquiry. In addition, we might ask how use of AI impacts users' mental health, frequency of interpersonal contact, beliefs about others' personalities and views, and other aspects of human life that are likely to improve or undermine interpersonal cooperation.
2. *What is the impact of value alignment strategies on interpersonal cooperation? Where there is conflict, how should we trade off the importance of value alignment against that of social alignment?* While I have not yet considered questions of this sort, the concerns about AI raised in Sect. 3 have natural application to some techniques of value alignment. As an example, consider RLHF, which aligns AI models through labor-intensive human evaluation of their responses. If an AI model with special prominence is aligned by using the feedback of humans who represent a subset of all human values and preferences, then the model may be aligned only with those values and preferences. Widespread use of the model may result in those values and preferences coming to be seen as more appropriate or legitimate, or it may cause tension between those who share those values and those who don't. As this example shows, the study of social alignment does not exclude the study of value alignment, and may even presuppose it. But social

alignment questions can highlight important blindspots in a pure value alignment approach. In particular, it can highlight open normative and conceptual questions that are not settled by even those engineering techniques that can provably align an AI agent.

I have presented some issues that should be on the social alignment agenda and are not currently studied as a matter of value alignment. But why think that value alignment leaves these issues out as a conceptual matter? After all, the idea of value alignment is capacious. Isn't social alignment simply value alignment but with interpersonal cooperation as a goal? There are two responses to this objection. The first response is that, when it comes to setting the AI safety research agenda, practical questions about what issues we should attend to are far more important than precisely defining the boundaries of our concepts. That is to say that we may well concede that social alignment is just a further specification of the concept of value alignment, yet point out that thus far the people and institutions pursuing value alignment have neglected the questions of social alignment, giving us pragmatic reason to assign them a distinctive label as a rallying cry for a new agenda.

The second response is that, as a conceptual matter, there is good reason to think that specifying value alignment with the goal of interpersonal cooperation still leaves out important aspects of the idea of social alignment. Recall that value alignment studies how AI systems can be engineered to produce consequences compatible with our goals. This leaves out those factors external to the AI system that determine the consequences it produces, in particular human and institutional responses to those systems. The social norms and networks involved in the conjectures of Sect. 3 are part of this social response to AI, and this is why they have been largely left out of the study of value alignment. We can therefore think of social alignment as an *ecological* agenda, one that attends not only to the engineering details of AI systems but also their operation within a given social environment, and that pays attention to the cumulative effects of AI systems and not just whether individual systems potentially cause harm.

In addition, the definition of value alignment assumes that we have goals that an AI system should respect. But cooperation plays an important role in the absence of fixed goals—in particular, when we form our goals, or allow our goals to change in response to experience and the needs of others, or specify our goals in relation to new knowledge or circumstances. Against this point the following objection may be raised: isn't cooperation a kind of procedural constraint or meta-goal to which we can apply value alignment methods? But this betrays a misunderstanding of the argument I have been articulating. Social alignment does not aim to maximize cooperation or even maintain the status quo. Rather, it is an investigation of the impact of AI upon the cooperative styles and infrastructure that make up the cooperative interface. How we evaluate this impact and respond to it are normative questions that should themselves be part of the study of social alignment, rather than simply assumptions fed into the engineering techniques of value alignment.

As an illustration of the objection that value alignment could encompass the concerns raised here, consider that there are already strands of the AI safety literature that explicitly study cooperation. The study of *legibility* concerns the ability of humans working with robots to predict their movements and understand their intentions (Lichtenthäler and Kirsch 2014; Dragan et al. 2013). This work facilitates better human-robot cooperation, itself an important theme in AI safety (Sandoval et al. 2016). Other developments in human-AI cooperation include the study of *Interpretable Machine Learning*, which seeks to explain the opaque predictions of machine learning models to their users (Molnar et al. 2021), and *affective computing*, which seeks to give computers the ability to engage with human emotions, either by reacting to or expressing them (Picard 1997; Robinson and El Kaliouby 2009).

Yet while these explorations may well illuminate important aspects of the cooperative interface, none of them directly study the cooperative interface itself. Each field explores the features and engineering of computing systems that meet certain constraints, rather than the way in which these systems affect interpersonal relations, let alone the conceptual and normative questions about the sorts of impact we should aim at. The closest we come to this is the research program on *computer-supported cooperative work* (CSCW), especially active in the 1990s, that considered how computers affect teamwork (Nass et al. 1996; Nass and Moon 2000), recommended the design of technology that could aid workplace collaboration (Greif 1988; Galegher et al. 1990; Olson and Olson 2007), and even warned about how ostensibly collaborative technologies would still benefit some within the organization and burden others (Grudin 1988).

The CSCW literature suggests a similar exploration of how other areas of computer technology, particularly AI, impact interpersonal cooperation. The social alignment agenda calls for reviving this line of research, and in particular the conceptual and normative questions regarding cooperation that arise in the context of AI. What do we want cooperation to look like in the age of intelligent machine agents, given how they affect the affordances for interpersonal cooperation? To return to Sect. 2's metaphor of driving: the study of human-robot cooperation is like learning about the car's interface—how to use the turning signals and so on; the study of AI social alignment is instead like learning about the cooperative interface made up by the roads, traffic laws and practices, and the cooperative affordances provided by car technology. The former sort of inquiry may be useful for the latter, but it cannot stand in for it.

I have been emphasizing the empirical questions that the social alignment agenda will have to address, and this is because the empirical questions help us to distinguish the social alignment agenda from the value alignment agenda by focusing our attention on the interaction of AI and our social institutions. But questions of social alignment are not purely empirical. Once we have a clearer idea of the way in which AI might impact the cooperative interface, we should expect that we will face tradeoffs that give rise to normative questions. For example, the impact of AI upon the pedagogical and productive functions of the workplace may well pull in different directions. In particular, using AI may help the productive function of the workplace by reducing the need for cooperation; but in doing so, it may decrease the opportunities for workers to practice cooperation with each other, and this may undermine the pedagogical function of the workplace. So we would be faced with a normative question: what balance should we want the workplace to strike between these functions? This would help answer the partly empirical and partly normative question:

how should we manage and regulate the automation of work? But we need further insight into how AI impacts the cooperative interface of the workplace before we can properly frame such normative questions.

Acknowledgements This paper benefited from discussion with participants at the 2022 Workshop on AI and the Future of Work at TU Delft. I am very grateful to two anonymous reviewers for their thorough and constructive recommendations, and to Filippo Santoni de Sio for his patience throughout the editorial process.

Funding No funds, grants, or other support was received. The author has no relevant financial or non-financial interests to disclose.

References

Agrawal, A., J. Gans, and A. Goldfarb. 2018. *Prediction machines: The simple economics of artificial intelligence*. Boston: Harvard Business Review Press.

Alchian, A.A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Almond, G.A., and S. Verba. 1963. *The civic culture: Political attitudes and democracy in five nations*. Princeton: Princeton University Press.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. 2016. Concrete problems in AI safety, *arXiv:1606.06565*.

Anderson, E. 2010. *The imperative of integration*. Princeton: Princeton University Press.

Anderson, E. 2017. *Private government*. Princeton: Princeton University Press.

Aquilina, M., E. Budish, and P. O'Neill. 2022. Quantifying the high-frequency trading "Arms Race." *Quarterly Journal of Economics* 137(1): 493–564.

Arora, S., and P. Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297: 103500.

Bai, Y.A., K. Jones, K. Ndousse, and 27 others. 2022. Training a helpful/harmless assistant with reinforcement learning from human feedback. *arXiv: 2204.05862*.

Banaji, M.R., and C.D. Hardin. 1996. Automatic stereotyping. *Psychological Science* 7(3): 136–141.

Bartneck, C., D. Kulić, E. Croft, and S. Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1: 71–81.

Bernstein, E.S., and S. Turban. 2018. The impact of the "Open" workspace on human collaboration. *Philosophical Transactions of the Royal Society B* 373: 20170239.

Bicchieri, C. 2006. *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bicchieri, C., and A. Lev-On. 2007. Computer-mediated communication and cooperation in social dilemmas: An experimental analysis. *Politics Philosophy & Economics* 6: 139–168.

Bobu, A., A. Peng, P. Agrawal, J. Shah, and A.D. Dragan. 2023. Aligning robot and human representations. *arXiv: 2302.01928*.

Bostrom, N. 2014. *Superintelligence*. Oxford: Oxford University Press.

Bratman, M.E. 1987. *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M.E. 1992. Shared cooperative activity. *Philosophical Review* 101: 327–341.

Bratman, M.E. 2014. *Shared agency*. Oxford: Oxford University Press.

Brosig, J., J. Weimann, and A. Ockenfels. 2003. The effect of communication media on cooperation. *German Economic Review* 4(2): 217–241.

Brown, M.A., J. Bisbee, A. Lai, R. Bonneau, J. Nagler, and J.A. Tucker. 2022. Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. Available at SSRN: https://ssrn.com/abstract=4114905

Burt, R.S. 1992. *Structural holes: The Social structure of competition*. Cambridge, MA: Harvard University Press.

Burt, R.S. 2005. *Brokerage and closure: An introduction to social capital*. Oxford: Oxford University Press.

Burt, R.S. 2010. *Neighbor networks: Competitive advantage local and personal*. Oxford: Oxford University Press.

Carlsmith, J. 2022. Is power-seeking AI an existential risk? *arXiv:2206.13353*.

Chen, A.Y., B. Nyhan, J. Reifler, R.E. Robertson, and C. Wilson. 2022. Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. *arXiv preprint* arXiv:2204.10921.

Chita-Tegmark, M., M. Lohani, and M. Scheutz. 2019. Gender effects in perceptions of robots and humans with varying emotional intelligence. In *14th ACM/IEEE international conference on human-robot interaction (HRI)*, 230–238. https://doi.org/10.1109/HRI.2019.8673222

Christian, B. 2020. *The alignment problem*. New York: W.W. Norton & Company.

Christiano, P.F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Cohen, M.K., B. Vellambi, and M. Hutter. 2020. Asymptotically unambitious artificial general intelligence. *arXiv:1905.12186.*

Crowell, C.R., M. Scheutz, P. Schermerhorn, and M. Villano. 2009. Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3735–3741.

Dasgupta, N. 2013. Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology* 47: 233–279.

Dragan, A., K. Lee, and S. Srinivasa. 2013. Legibility and predictability of robot motion. In *ACM/IEEE international conference on human-robot interaction*, 301–308.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Erickson, T., and D.W. McDonald. 2007. *HCI remixed: Rreflections on the works that have influenced the HCI community*. Cambridge, MA: MIT Press.

Estlund, C. 2003. *Working together*. Oxford: Oxford University Press.

Everitt, T., M. Hutter, R. Kumar, and V. Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese* 198: 6435–6467.

Eyssel, F., and F. Hegel. 2012. (S)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology* 42: 2213–2230.

Ford, H. 1922. *My life and work, in collaboration with S. Crowther*. Garden City, NY: Garden City.

Frega, R., L. Herzog, and C. Neuhäuser. 2019. Workplace democracy–the recent debate. *Philosophy Compass* 14(4): 1–11.

Frey, C.B., and M.A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting & Social Change* 114: 254–280.

Fullwood, C., S. Quinn, L.K. Kaye, and C. Redding. 2017. My virtual friend: A qualitative analysis of the attitudes and experiences of smartphone users: Implications for smartphone attachment. *Computers in Human Behavior* 75: 347–355.

Gabriel, I., and V. Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to AI safety. In *The Oxford Handbook of digital ethics*, Online Ed., ed. C. Veliz. Oxford: Oxford University Press.

Galegher, J., R.E. Kraut, and C. Egido. 1990. *Intellectual teamwork: Social and technological foundations of cooperative work*. New York: Taylor & Francis.

Ganguli, D., A. Askell, N. Schiefer, T.I. Liao, and K. Lukošiūtė, and 44 others. 2023. The capacity for moral self-correction in large language models. *arXiv: 2302.07459*.

Granovetter, M.S. 1973. The strength of weak ties. *American Journal of Sociology* 78: 1360–1380.

Granovetter, M.S. 1995. *Getting a job: A study of careers and contacts*, Second Edition. Chicago: University of Chicago Press.

Greenwald, A.G., D.E. McGhee, and J.L. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74(6): 1464–1480.

Greif, I. 1988. *Computer-supported cooperative work: A book of readings*. San Mateo: Morgan Kaufmann.

Grudin, J. 1988. Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proceedings of the 1988 ACM conference on computer-supported cooperative work,* 85–93.

Hadfield-Mennell, D., A. Dragan, P. Abbeel, and S. Russell. 2016. Cooperative inverse reinforcement learning. *arXiv: 1606.13137.*

Hadfield-Mennell, D., A. Dragan, P. Abbeel, and S. Russell. 2017. The off-switch game. *arXiv:1611.08219*.

Hibbard, B. 2012. Model-based utility functions. *Journal of Artificial General Intelligence* 3: 1–24.

Holtman, K. 2019. Corrigibility with utility preservation. *arXiv:1908.01695*.

Hsieh, N. 2005. Rawlsian justice and workplace republicanism. *Social Theory and Practice* 31(1): 115–142.

Huebner, B. 2016. Implicit bias, reinforcement learning, and scaffolded moral cognition. In *Implicit bias and philosophy Vol. 1*, eds. M. Brownstein, and J. Saul, 47–79. Oxford: Oxford University Press.

Hwang, T. 2020. *Subprime attention crisis: Advertising and the time bomb at the heart of the internet*. New York: Farrar, Straus and Giroux.

Jensen, C., S.D. Farnham, S.M. Drucker, and P. Kollock. 2000. The effect of communication modality on cooperation in online environments. *CHI Letters* 2000: 470–477.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Kearns, M., and A. Roth. 2020. *The ethical algorithm: The science of socially aware algorithm design*. New York: Oxford University Press.

Kiesler, S., L. Sproull, and K. Waters. 1996. A prisoner's Dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology* 70: 47–65.

Klein, B. 1988. Vertical integration as organizational ownership: The fisher body-general motors relationship revisited. *Journal of Law Economics and Organization* 4(1): 199–213.

Knorr Cetina, K. 2009. The synthetic situation: Interactionism for a global world. *Symbolic Interaction* 32(1): 61–87.

Knorr Cetina, K., and U. Bruegger. 2002. Traders' engagement with markets: A postsocial relationship. *Theory Culture & Society* 19: 161–185.

Krakovna, V., and K. János. 2023. Power-seeking can be probable and predictive for trained agents. *arXiv: 2304.06528.*

Kuchenbrandt, D., M. Häring, J. Eichberg, F. Eyssel, and E. André. 2014. Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *International Journal of Social Robotics* 6: 417–427.

Kumar, R., J. Uesato, R. Ngo, T. Everitt, V. Krakovna, and S. Legg. 2020. REALab: An embedded perspective on tampering. *arXiv: 2011.08820*.

Lange, A.-C. 2016. Organizational ignorance: An ethnographic study of high-frequency trading. *Economy and Society* 45(2): 230–250.

Leike, J., D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. 2018. Scalable agent alignment via reward modeling: A research direction. *arXiv* 181107871.

Lewis, M. 2014. *Flash boys: A wall street revolt*. New York: W.W. Norton & Co.

Lichtenthäler, C., and A. Kirsch. 2014. Goal-predictability vs trajectory-predictability: Which legibility factor counts. In *2014 ACM/IEEE international conference on human-robot interaction*, 228–229.

Lorenz-Spreen, P., L. Oswald, S. Lewandowsky, and R. Hertwig. 2022. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behavior*. https://doi.org/10.1038/s41562-022-01460-1.

MacDonald, S., N. Lin, and D. Ao. 2009. Networks of opportunity: Gender, race, and job leads. *Social Problems* 56(3): 385–402.

MacKenzie, D. 2015. Mechanizing the Merc: The Chicago mercantile exchange and the rise of high-frequency trading. *Technology and Culture* 56(3): 646–675.

MacKenzie, D. 2017. A material political economy: Automated trading desk and price prediction in high-frequency trading. *Social Studies of Finance* 47(2): 172–194.

MacKenzie, D. 2018. Material signals: A historical sociology of high-frequency trading. *American Journal of Sociology* 123(6): 1635–1683.

MacKenzie, D. 2019. How algorithms interact: Goffman's "Interaction Order" in automated trading. *Theory Culture & Society* 36(2): 39–59.

MacKenzie, D. 2021. *Trading at the speed of light*. Princeton: Princeton University Press.

Marchesi, S., D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology* 10(450): 1–13.

Melamed, L. 2009. *For crying out loud: From open outcry to the electronic screen*. Hoboken: John Wiley & Sons.

Melumad, S., and M.T. Pham. 2020. The smartphone as a pacifying technology. *Journal of Consumer Research* 47: 237–255.

Melumad, S., and R. Meyer. 2020. Full disclosure: How smartphones enhance consumer self-disclosure. *Journal of Marketing* 84: 28–45.

Molnar, C., G. Casalicchio, and B. Bischl. 2021. Interpretable machine learning—A brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases, in communications in computer and information science*, Vol 1323, 417–431.

Narveson, J. 2001. *The libertarian idea*. Guelph: Broadview.

Nass, C., and Y. Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56: 81–103.

Nass, C., B.J. Fogg, and Y. Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45: 669–678.

Ng, A.Y., and S.J. Russell. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning*, 663–670.

Olson, G.M., and J.S. Olson. 2007. Groupware and computer-supported cooperative work. In *The Human-computer interaction handbook*, 2nd edition, ed. A. Sears and J.A. Jacko, 545–558. Boca Raton: CRC Press.

Patterson, S. 2012. *Dark pools: The rise of the machine traders and the rigging of the U.S. Stock Market*. New York: Random House.

Pedulla, D.S., and D. Pager. 2019. Race and networks in the job search process. *American Sociological Review* 84(6): 983–1012.

Picard, R.W. 1997. *Affective computing*. Cambridge, MA: MIT Press.

Powers, A., A.D.I. Kramer, S. Lim, J. Kuo, S.-L. Lee, and S. Kiesler. 2005. Eliciting information from people with a gendered humanoid robot. In *IEEE International Workshop on Robot and Human Interactive Communication,* 158–163.

Preda, A. 2013. Tags, transaction types and communication in online anonymous markets. *Socio-Economic Review* 11: 31–56.

Rajan, R.G. 2006. Has finance made the world riskier? *European Financial Management* 12(4): 499–533.

Rajan, R.G., and L. Zingales. 2001. The firm as a dedicated hierarchy: A theory of the origins and growth of firms. *Quarterly Journal of Economics* 116(3): 805–851.

Rawls, J. 1999. *A theory of justice*. Revised ed. Cambridge, MA: Harvard University Press.

Robinson, P., and R. Kaliouby. 2009. Computation of emotions in man and machines. *Philosophical Transactions of the Royal Society B*. 364: 3441–3447.

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. New York: Viking.

Sandoval, E.B., J. Brandstetter, M. Obaid, and C. Bartneck. 2016. Reciprocity in human-robot interaction. *International Journal of Social Robotics* 8: 303–317.

Satz, D. 2023. What is wrong with the commodification of human labor power: the argument from "democratic character". In *Working as equals: Relational egalitarianism and the workplace*, eds. J.D. Jonker, and G. Rozeboom. Oxford: Oxford University Press.

Siegel, M.C., Breazeal, and M.I. Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2563–2568.

Skalse, J., N.H.R. Howe, D. Krasheninnikov, and D. Krueger. 2022. Defining and characterizing reward hacking. *arXiv:2209.13085*.

Soares, N. 2016. The value learning problem. In *Ethic for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence*, 9–15 July. https://intelligence.org/files/ValueLearningProblem.pdf

Soares, N., B. Fallenstein, E. Yudkowsky, and S. Armstrong. 2015. Corrigibility. In *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, 25–26 January 2015. https://intelligence.org/files/Corrigibility.pdf

Suchman, L.A. 2012. *Human-machine reconfigurations*. 2nd ed. Cambridge: Cambridge University Press.

Susskind, D. 2020. *A world without work*. New York: Henry Holt.

Tay, B., Y. Jung, and T. Park. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* 38: 75–84.

Taylor, J., E. Yudkowsky, P. LaVictoire, and A. Critch. 2020. Alignment for advanced machine learning systems. In *Ethics of artificial intelligence*, ed. S. M. Liao. New York: Oxford University Press.

Tsuruda, S. 2020. Working as equal moral agents. *Legal Theory* 26(4): 305–337.

Tufekci, Z. 2017. *Twitter and tear gas: The power and fragility of networked protest*. New Haven: Yale University Press.

Tufekci, Z. 2018. Youtube, the great radicalizer. *New York Times*, March 10, 2018, , 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtubepolitics-radical.html

Walker, C. R., and R. H. Guest. 1952. *The man on the assembly line*. Cambridge, MA: Harvard University Press.

Williamson, O. 1973. Markets and hierarchies: Some elementary considerations. *American Economic Review* 63: 316–325.

Williamson, O. 2010. Transaction cost economics: The natural progression. *American Economic Review* 100: 673–690.

Wängberg, T., M. Böörs, E. Catt, T. Everitt, and M. Hutter. 2017. A game-theoretic analysis of the off-switch game. *arXiv: 1708.03871*.

Yellen, J. L. 1984. Efficiency wage models of unemployment. *American Economic Review* 74(2): 200–205.

Zaharudin, K.Z., and M.R. Young, W.-H. Hsu. 2021. High-frequency trading: Definition, implications, and controversies. *Journal of Economic Surveys* 36: 75–107.

Zaloom, C. 2006. *Out of the pits: Traders and technology from Chicago to London*. Chicago: University of Chicago Press.

Złotowski, J., D. Proudfoot, K. Yogeeswaran, and C. Bartneck. 2015. Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics* 7: 347–360.