EPIC2019

Ethnographic Praxis in Industry Conference Proceedings

Calibrating Agency

Human-Autonomy Teaming and the Future of Work amid Highly Automated Systems

LAURA CESAFSKY, *Alliance Innovation Lab – Silicon Valley*
ERIK STAYTON, *Alliance Innovation Lab – Silicon Valley, MIT*
MELISSA CEFKIN, *Alliance Innovation Lab – Silicon Valley*

This paper explores how the design of everyday interactions with artificial intelligence in work systems relates to broader issues of interest to social scientists and ethicists: namely human well-being and social inequality. The paper uses experience designing human interactions with highly automated systems as a lens for looking at the social implications of work design, and argues that what human and automation each do is less important than how human and automation are structured to interact. The Human-Autonomy Teaming (HAT) *paradigm, explored in the paper, has been a promising alternative way to think about human interactions with automation in our laboratory's research and development work. We argue that the notion of teaming is particularly useful in that it encourages designers to consider human well-being as central to the operational success of the overall human-machine system that is being designed.*

> To think in interaction with a computer in the same way that you think with a colleague whose competence supplements your own will require much tighter coupling between man [sic] and machine … than is possible today.
>
> - J. C. R. Licklider, "Man-Computer Symbiosis" (1960)

INVENTING AND CALIBRATING A HUMAN-AUTONOMY TEAM

An operator sits in front of a giant, curved monitor on an otherwise Spartan white desk. With mouse and keyboard, she interacts remotely with an autonomous vehicle (AV) out on the roadway that needs, and has 'called for,' her help. The AV 'wants' to go around an obstacle—a double-parked delivery vehicle—that impedes its progress, but it is not sure if it should. The young woman clicks a series of buttons and, in response to her input, the car cautiously edges out, crosses the double yellow line, and drives around the obstruction to continue on its journey. This action may not seem like much, but our operator has just engaged in a delicate ballet of Human-Autonomy Teaming ("HAT" for short).

This paper explores how the design of these everyday interactions with artificial intelligence in advanced work systems might relate to broader issues of interest to social scientists and ethicists working in technology, such as human well-being and social inequality. It draws ethnographically on our experiences working intensively with engineers in an AV Innovation Lab to design how agency in collective problem solving will be distributed across human and non-human agents in SAM, our Seamless Autonomous Mobility system. SAM supports the remote management of fleets of AVs in times of trouble; one of its chief value-adds is the ability to bring human intelligence into an otherwise-automated technical loop in crucial moments. Yet exactly how, when, and why the intelligence of this "Mobility Manager" should be engaged via SAM has been the subject of

intense speculation, experimentation and debate among the multi-disciplinary researchers in our lab. We refer to this contested process as *calibrating agency*.

Our essay is situated amid this contest over the proper calibration of the Mobility Manager's agency in the SAM system, and in response to a growing body of scholarship at EPIC and elsewhere on automation, human work, and the 'end of the job' as we know it (e.g. Cefkin et al 2014; Yapchaian 2018). Our contribution is to offer new insights on the topic of meaningful work in relation to current debates about automation.

Our first overarching theme is that we should not associate automation only with humans being tossed "out" of the technical loop at work (Bradshaw et al 2013; Gray & Suri 2019). As serious as the issue of worker displacement is, in our work we have experienced the other side of the coin: that purportedly automated technologies like AVs do in fact need human workers and their human agencies "on the team" and "in the loop" during real-time operations in order to function. The growing need to invent jobs for technologies formerly thought to be "automated" presents a practical and intellectual opportunity for ethnographers and others working in technology to attempt to influence the automation process towards more humane outcomes. To succeed at this task we will need hybrids: of human and machine, of research and design, and of academic and applied sensibilities (Blomberg 2005).

At work in our lab, as we detail in the body of the paper, the question of what the Mobility Manager ought to do to help in AV problem-solving has often been figured in terms of "role" rather than automation paradigm—that is in terms of *what* rather than *how*. Should the operator be deployed as a "social-knower"; a "technical band-aid"; an "AI machine trainer'; a "legal actor"? The paper uses the example of the Obstructions use case for SAM (which appeared in our opening vignette) and these four different roles as a way to unpack what Teaming means in terms of 'how' the operator is imbricated in highly automated systems, and the challenges that different paradigms raise for worker well-being.

To that end, the second overarching theme of the paper is exploring Human-Autonomy Teaming as an emerging automation paradigm, and a framework for designing the SAM operator work role toward more ethical outcomes. HAT is a human-machine interaction paradigm focused on creating reliable and efficient interfaces for managing human-autonomy interactions in safety-critical decision-making systems. Yet we describe how the optimistic ethos of Teaming—'bring the best out of each teammate, human and machine alike!'—leaves practical space to research and advocate for operator workflows that consider issues like worker alienation, culpability for system error, and the growing rift between 'haves' and 'have-nots' in high-technology economies. Dealing with these problems, we will suggest, means not simply giving the worker "more" agency with respect to the machine, but instead attending to the intricate details of implementing collaboration.

HISTORIES OF AGENCY IN THE DESIGN OF LABOR

While the business rhetoric around AI, machine learning, and predictive analytics argues that human beings can be *eliminated* from a wider and wider range of tasks—and profits and user outcomes thereby improved—, we know from a long history of automation studies that the reality is never this simple. Human roles and agencies are displaced, shifted in time and space, but not simply eliminated or made obsolete (Mindell 2015).

Labor automation is at least as old as the wind and water mills of the Middle Ages. Something close to the modern rhetoric of high automation can be found already in Oliver Evans's "fully automated" grist mills of the 1780s: romanticized descriptions of this mechanized production line for grain consistently downplayed the roles of human tenders in management, maintenance, and implementation (White 1962; Smith 2016). Taylorism and the assembly line are the better-known successors to the Evans Mill, and made more explicit the roles of the human being within the automated system: to be part of the machine oneself, performing a rote labor process in a precisely choreographed way; or to be a machine engineer, ensuring the automation does its job and carrying out via machine technology strategic decision-making tasks (Taylor 1911; Diebold 1959; Aitken 1960).

Cynically, then, automation has two different valences, from two different subject positions. For some, what it means to have agency in an increasingly automated world is to be a human body that is itself a tool of technology: instead of technologies being 'mediator[s] between man and the world', humans become mediators between technology and the world (Simondon 2011). Such is the world of the machine tender. For others, agency is increasingly expressed by wielding machines: designing them, ordering them about, and using them (along with their associated human tools) to free up more time and energy for creative work (Noble 1984). Such is the world of the engineer or manager. This dance of "managerial" and "shop-floor" agencies, mixed in with the agencies of machines, continues everywhere from Shenzhen to the surface of Mars. We see it show up again, as we explore in the body of the paper, in contemporary automation paradigms like microwork and supervisory control that are proposed for real-time oversight of 'autonomous' systems.

HAT inserts itself into this "master-slave" dualism, where one is either ruled by or rules the machine, with the dreamy-eyed proposition that the most effective way to enmesh humans and AI is to make them equals of sorts—to "team" them. HAT is therefore the spiritual successor to J. C. R. Licklider's 1960 vision of human-machine symbiosis (Licklider 1960). HAT emerged from human-machine interaction literature, and especially from research in the aviation domain, as a field of technical specialty. It makes the argument that, especially given the complex domains in which automated technologies aspire to operate today, outcomes are less effective when human operators have either too much or too little agency, or when automation relationships are rigid, as with Taylor's assembly lines (Brandt et al 2017; Endsley 2017). Teaming tries to retains the benefits of automation—mainly, efficiency—while minimizing two of its chief costs and hazards—especially brittleness (the inability to adapt to new situations and contexts) and alienation of the operator (Shively et al 2017). The promise is that, on a team, neither humans nor technology become the tool: rather, they work together creatively to solve increasingly complex problems.

Behind this optimistic rhetoric lie sober research problems that AI and HMI researchers are just beginning to tackle. There are very many practical and technical questions of team-building, and a growing research agenda on the philosophical and pragmatic implications of machines as teammates—both from the robot and the human ends (Schaefer et al 2017). After all, effective teamwork is an intricate engineering challenge that requires generating "actual coordination of complex activities such as communication, joint action, and human-aware execution to successfully complete a task, with potentially shifting goals, in varying environmental conditions mired in uncertainty" (Seeber et al 2019, 3). Because with HAT neither roles nor tasks are defined in advance, and because finding the optimal form of

'teamwork' is an experimental problem unique to each system, HAT affords—and requires—a more inventive *calibration of agency* than other automation paradigms.

The definition of agency is a central, contested concept in philosophy. In deploying the term we risk entering relentlessly muddy waters, and do not seek to resolve the contest. Rather, we endeavor to define only what *we* mean practically when we speak of agency in work design and in relation to the machine. We forward a minimalist definition of agency, drawn from Actor-Network Theory and science studies: agency as the simple capacity—shared by humans and non-humans alike—to alter the course of events in some situation. Agency can be recognized by asking the following of an entity: "[d]oes it make a difference in the course of some other agent's action or not? Is there some trial that allows someone to detect this difference?" (Latour 2005, 71). We might also glean this in the reverse: if the human is inserted into the loop of automation only to supervise or "rubber stamp" automated processes that would have unfolded exactly the same way in their absence, then we can conclude that they are not exercising agency. This definition of agency thus does not say anything specific about the concerns of the classic philosophers of agency—the more humanistic visions that worry about the place of human will, intentionality, reason, and self-realization (Kockelman 2013). Yet we do reunite with that tradition in a more obtuse way, in the sense that we are interested in how automation paradigms like HAT might produce more engaging and reasonably remunerated jobs that might allow a worker to lead a dignified life and, to the extent possible, influence the direction and possibilities in her life.

OUR WORK AT THE INNOVATION LAB

The automobile industry is by outward appearances a paradigmatic case of the automation of human labor out of an existing system. Indeed, in the earlier days of the AV industry, many of the bigger players operated under the assumption that the software would entirely replace human oversight (Markoff 2014). However, as technological setbacks have sobered the industry, this attitude has shifted, and exploring human-in-the-loop technology has become *de rigueur* (Harris 2018; Davies 2018). History shows us that this should be no surprise: technologies that are autonomous inside the lab regularly involve humans-in-the-loop by the time they leave it. Examples from spaceflight have shown the continued need to involve human judgment and flexibility, whether in person (Mindell 2011) or at a distance (Clancey 2014).

As researchers at a major manufacturer's AV Innovation Lab, our everyday work is mostly about how to keep various humans "in the loop": aware of, in-step with, and in seamless and positive interaction with the purportedly "autonomous" vehicle systems we are creating. Especially driven by the director of Nissan Research in Silicon Valley at the time, and the principal scientists for Autonomous Vehicle development—both of whom had come from NASA—our AV lab was perhaps unique in that there was an early and strong belief that autonomous systems would always need humans in the loop somewhere. Or, at the very least, they would be needed for quite some time to accelerate the process of getting AV on the road. The Seamless Autonomous Mobility (SAM) human-in-the-loop vehicle management system has been one of the main research efforts at the Lab from its opening in Silicon Valley in 2013, and it was constructed around that same intuition.

Yet this conviction that a human-in-the-loop is necessary was, and still is, an article of faith first. What exactly humans are needed in the loop to do remains an object of

considerable debate. There is a gestalt sense that we need humans to make automation work, but debate and shifting positions over the projected capabilities of machines—and therefore controversy over the required roles for humans in automated systems—is abundant. This is in part because of uncertainty about what AI will and will not be capable of in the future: it is a 'teammate' whose future skills we can only guess.

As social scientists designing for the roles of these humans within the vehicle system, we have been in the thick of things as active participants in the *interessement* and *enrolment* of actors into these sociotechnical visions (Callon 1984). We have been working closely with multi-disciplinary teams of engineers and designers for several years to create a work role for the Mobility Manager within the SAM system. This work has involved studying analog fleet management roles in aviation and public transportation. It also involved studying real-world use cases for SAM where the insertion of human agency into an automated loop is likely to be vital, now and in the future. This year we have moved from research to the design phase, taking a leadership role in the creation of experimental systems for effective collaboration between humans and autonomy. We are currently collaborating on building a prototype of a front-end teaming interface and back-end teaming manager for SAM.

Our work on this experimental prototype has been influenced and aided by a collaboration we established with a team of Human Systems Integration researchers at NASA's Ames Research Center who study the future role of automation in national airspace management. They have been working on validating a HAT paradigm that seeks to find a 'sweet spot' between too much human labor and too much brittle and alienating automation.

Their approach to interface and system design emphasizes a few key principles which we will explore further in the next section: 1) careful provision of information to support full situational awareness for the operator 2) transparency to allow the operator to understand of what automation is doing and how they can affect its actions; 3) bi-directional communication to allow human and AI to work collaboratively to generate and evaluate options and make decisions; 4) variable levels of automation (LOAs) that put neither human nor automation exclusively in charge of most tasks; and 5) a "playbook" concept that brings it all together, wherein collaborative action is enacted quickly by predefined scenarios at variable LOAs with set goals, roles, and responsibilities, and that the human and the autonomy settle-on collaboratively in response to different real-world scenarios they face (Brandt et al. 2017).

By experimenting with these principles in our work, we aim to make Mobility Management not only efficient and safe, but also ethical and engaging, as we incorporate new capabilities our engineers are developing for our AI 'teammates', such as robot introspection and self-explanation. As practitioners in industry we must remain focused on efficiency and functional fleet management foremost. Yet teaming's feel-good ethos of 'bringing out the best in everyone,' and its promise of flexibility in designing interactional relationships, leaves room to stretch out into implications for ethical and political domains—especially since design prescriptions such as "transparency," as we will see, operate deeply on both the functional and ethical levels. This has left us room to more quietly address issues brought up in the work of anthropologists of technological labor (Gray and Suri 2019; Elish 2019), especially the ethical consequences of calibrating agency.

DISCUSSION: HUMAN-AUTONOMY TEAMING IN ACTION

The HAT prototype we are currently designing began with the management of human involvement in just one type of on-road use case: the 'obstructions' case described briefly in the introduction. Obstructions are an early and paradigmatic case for the use of a human-in-the-loop in AV systems, and elaborating them here provides a good example of the contest over human role and function and the kind of 'cut' that HAT takes at the question. It exposes connections between how "micro-interactions" between AI and operators are implemented, and ethical and "macro-" consequences for three domains: worker alienation; growing economic inequality; and worker culpability for accidents.

Obstructions cases are usually easy situations for human drivers to handle, so easy we do them without conscious thought. You see a delivery truck parked in your driving lane with its flashers on, and quickly you do a number of things: determine if it is legitimate to try to overtake it; determine if it is safe to overtake it, even though you have to cross momentarily into the other traffic lane; and initiate a way to overtake it.

But obstructions are actually quite difficult for autonomous systems to handle on their own today, for reasons that are being actively researched and debated. Each of these reasons might be understood as a potential opportunity for teamwork and a "role" for the Mobility Manager: as "social-knower;" "technical band-aid;" AI "machine trainer;" and "legal actor" as described below. These positions are not mutually exclusive. And each of these positions has had, at different moments, different supporters among key technologists and decision makers in the lab, who grapple over which parts and capabilities of the human operator to make use of in order to divergent technological and business goals.

We as UX researchers, at least ideally, represent the interests of the human— rather than technological, business or other kinds of interests—in the design process. Looking at this internal debate among stakeholders about the human's role in the system, it becomes pertinent to ask: "What is in the interest of the human being with respect to these types of potential roles within complex, multi-agent systems?"

Contending Work Roles

Human as "Technical Band-Aid"

> *"[The] vehicle can tell the traffic state, and even recognize some hand gestures, but human judgment is required for the appropriate course … The request is routed to the mobility manager, who decides on the correct action, and creates a safe path around the obstruction"* (Nissan 2017a)

The earliest technical capacity given to the human in SAM was *teleoperation*: the ability to direct an AV along a human-drawn path forward, not by remotely driving (or "joysticking") the car, but by sending it instructions (speed and directionality).[1] This capacity was useful for situations where the AV's ability to plan its own path was comprised, and was built upon NASA technology used to direct robots around the surface of Mars. Thus the first concept of a role for the human-in-the-loop made her into a technical band-aid, an agent that would make up for technical deficiencies with respect to *how* to go around an Obstruction. Teleoperations takes some risk out of the job of mobility management, as the AV always decides for itself *when* to go or stop and keeps its basic sensors and crash-avoidance

functions engaged. But this role does imply a human making up for technological lacks such as visibilities and insufficient maps, in an effort to streamline the development process and make possible early introduction of AVs.

Human as "Social Knower"

> *"How are you gonna know if you can go around? What's this guy waving trying to tell you to do? We will need a human to understand the situation and make that call."*
> – An employee in the laboratory, talking about SAM

Studying on-road Obstructions use cases, however, it was soon realized that the question of *if* an AV should go around an obstacle might be the bigger problem than figuring out *how* to go around one—especially as the technology improves and the need for technical band-aids decreases. Indeed, in more recent implementations of the Obstructions use case, the autonomy proposes its own path around the obstacle for most situations, and the human's role is simply to confirm or deny the social legitimacy of the maneuver.

The social knower vision is all about context. Understanding context in human terms and engaging fluently in the social domain have been longtime problems for automated systems. Treating the human mobility manager as a "social knower" is sometimes a pragmatic response to current difficulties, but it can also represent a broader philosophical position about the limits of AI, and the indelible place for the human in knowing specifically "social" or "human" things like the context of the situation (Is this really a passable object? Is that a cop directing me to go around, or just a person waving?). In this imagination, the human mobility manager is a contextual interpreter, a common-sense reasoner, and an indelible aspect of a successful system. Some managers at the lab have championed this role as the *raison d'être* of the Mobility Manager position.

Human as "Machine Trainer"

> *"The system learns and shares the new information created by the Mobility Manager. Once the solution is found, it's sent to the other vehicles. As the system learns from experience, and autonomous technology improves, vehicles will require less assistance and each mobility manager will be able to guide a large number of vehicles simultaneously."* (Nissan 2017b)

As the SAM system has further evolved, more attention has been given to how the system will improve over time. We do not want to just solve the case at hand, but get better at solving other similar cases. In this vision, the mobility manager is an annotator who is creating the data set that will allow a future AI to succeed where current AI has failed: labeling misrecognized objects in a scene, or modeling "good driving behavior" so that it can be copied. This vision is about machine learning. Spurred by advances in supervised machine learning via neural networks, there is great hope that, with enough labeled data, a clever architecture can solve any problem. But data *is* the problem. In a space as complex as that of the roadway—even just for obstacle avoidance scenarios—the number of examples needed might exceed tens of millions. In this view of the human's role, there are no philosophical reservations about unique human capabilities; she is just there to produce the necessary data.

Over time, her role becomes less and less necessary until perhaps she could be eliminated entirely by AI trained upon her own labor.

Human as "Legal Actor"

> *"What is the system going to do when it has to break rules? Are you going to allow it break rules? But how are you going to define what rule it can break when and how?"*
> – The Lab's Chief Technical Director, quoted in an interview (Margeit 2019)

This vision is about responsibility. Anyone who goes through driver training in the United States—and many people who get ticketed by law enforcement—can recognize the extent to which the legal rules and social norms of the roadway come into conflict. It is generally illegal, for example, to cross a double-yellow line in the US. It is also illegal to double-park one's vehicle in a travel lane. And the California Vehicle Code makes no exception to the line-crossing rule in this case. But if AVs cannot break the law sometimes to overtake illegally stopped vehicles, they will be largely incompatible with existing streets and human behaviors, something legal experts themselves have been recognizing (Law Commission 2018). In conflicts between multiple laws, or between laws and norms, this position on the mobility manager's role suggests that they will certify these normal, tacitly legal maneuvers such as permitting a vehicle to cross over a double yellow line to avoid an obstacle, when it is safe to do so. But, unfortunately, this kind of mobility manager could also be a scapegoat for the vehicle operator to offload responsibility in the event of an accident or citation from law enforcement.

The Social Costs of Work Roles

Role is helpful because it identifies where the AI is 'weak' and where humans are 'strong,' and therefore highlights use cases and reasons for including humans as teammates. Yet we have come to believe via our research on SAM and study of HAT that focusing on role alone is actually the wrong frame if we want to understand the ethical consequences of human-automation relationships. When evaluating work roles in light of ethical concerns, in other words, it may matter less *what* the Mobility Manager does—that is, their role as social knower or legal actor—and indeed they will likely occupy multiple of these roles at different times as they solve problems. Rather, what emerges as of more concern is the *how* of that function—the implementation of the interaction design, which may or may not have a direct relationship to imagined role.

In other words, what must be considered is the *automation paradigm* (Endsley 2017): the high-level model of how the human and automation will interact, how responsibilities will be allocated between them, and how these allocations will change in the course of operation. There are, as we will see, multiple ways that a social knower role, for example, might be implemented, from paradigms that literally take the conscious decision-making out of the process, to ones that put the human into a (troubled) supervisor position with respect to the autonomy. Each of these positions could be made part of a human-autonomy team picture, but each has often been envisioned in the Lab outside of the team frame, instead in ones that reproduce master-slave dynamics, such as microwork, supervisory control, and

engineering paradigms. Each automation paradigm raises technical issues and presents political and ethical consequences for the worker.

Worker Alienation

In order to illustrate an extreme case of what it might mean to produce the Mobility Manager within SAM as an alienated laborer, we turn to a series of discussions we were involved in during early 2019. A novel paradigm was proposed with a novel technology attached: brain-machine interfaces that can interpret pre-cognitive signals from human brains. The brain-machine interface—a helmet with sensors for brain activity—was imagined as a partial solution to the Obstructions use case, in that its wearer could generate quick "go or no-go" decisions when the time was right for a supervised AV to overtake an obstacle on the road. Such a scheme puts the human in the social knower role, but as a pre-cognitive "social reactor" responding based on instinct to live video of the scene.

This is an automation paradigm best described as *microwork* (Lehdonvirta 2016). Microwork, or micro-tasking, is an increasingly common automation paradigm that forms the basis of so-called "flexible work platforms" like Amazon's Mechanical Turk and Facebook's Content Moderator work regimes. Microwork is considered the smallest unit of work in a virtual assembly line, describing tasks for which no efficient algorithm has *yet* been devised, and that today require human intelligence to complete reliably (Irani 2015). Tasks like supervising an autonomous vehicle around an obstacle can be further chopped into these 'micro' subtasks, including image identification, transcription and annotation; content moderation; data collection and processing; audio and video transcription; and translation. The very Tayloristic idea here is that the proper way to insert human agency into the loop of AI is to define precisely the tiny inputs an operator will contribute to process.

Microtasks tend to be repetitive, menial and tedious—the kind of job it is easy to create, but not necessarily the kind of job that the creators would want for themselves. Microwork-intensive automation paradigms have the potential to alienate the worker from the experiences that, research shows, make work satisfying: doing a variety of kinds of tasks, using higher order processing and troubleshooting skills, managing situations, communicating with others, helping people, using creativity, learning and growing, and making independent decisions (Manyika et al. 2017). These are the kinds of things that, taken together, produce a profession or a craft rather than a menial job, and that give us the opportunity to connect and use our human capacities.

Like Marx, we are concerned with the degree to which a job allows one to express fundamental parts of one's humanity, or whether it suppresses those human aspects for the goal of efficiency or some other value. Marx wrote of alienation in these terms:

> It is true that labour produces marvels for the rich, but it produces privation for the worker. It produces palaces, but hovels for the worker. It procures beauty, but deformity for the worker. It replaces labour by machines, but it casts some of the workers back into barbarous forms of labour and turns others into machines. It produces intelligence, but it produces idiocy and cretinism for the worker. (Marx 1844)

While Marx was describing the conditions of workers in the 19th Century, such lines could just as easily describe a 'brain helmet job' working amid a 21st century, mostly-

autonomous technology. The political and ethical questions with microwork today are much the same as with assembly-line work, leading Horton (2011) in *Economics Letters*, referencing Marx's co-author Engels, to inquire into what he cleverly calls the "Condition of the Turking Class." The vision here is of workers doing the same tiny task over and over and over again, the value of the human whittled down to just one tiny capability. "Go, go, no go, go, no go"—read off brain signals.

Any of the above roles can ostensibly be turned into a microwork job—all it requires is the extreme limitation through interface and work design of the scope and variety of the human's agential contribution. Teaming, taken seriously, rules out microwork as a desirable human-machine future, and therefore presents a possible (if only inadvertent) wedge to the plight of the Turking Class. This is due, in particular, to its organizing concern with the perils of over-automation and brittleness, and the resulting emphasis on ensuring both situational awareness and meaningful decision-making on the part of the human actor. Particularly important for HAT is minimizing "confirmation bias," or the tendency of humans within highly automated decision-making systems to agree without really *thinking* with the AI's reading of situation and its plans (Endsley 2017).

From a HAT point of view, if the problem with machines is that they are brittle—unable to respond appropriately when the situational context in which they are acting shifts—then an enduring task for humans on teams is likely to be in helping the machines react dynamically to *novel* situations. And this means that rather that inputting the same datum the same way over and over, part of the human operator's job description should be to make holistic situational assessments, at least in some cases, and to have a latitude for creative response. Doing so requires providing the operator with full situational awareness—something that microwork and chunking deliberately deny. In the best HAT arrangements, a remote operator achieves situational awareness of the external environment *and* of the automation itself at the highest level: they know what is going on, what that means, and what may happen next, for both the internals of the system and the real-world outside (ibid).

A second area where HAT might intrinsically help is that, due to its emphasis on variable levels of automation, it might produce more variety on the job. An operator can 'call' plays at the highest level and dynamically adjust Levels of Automation for tasks and subtasks within a play based on contextual factors. For instance, if the operator is being overloaded by too many issues, they can potentially: allow the automation to take full control of the least critical cases; check AI's suggestions for medium-risk cases; and be themselves totally in charge of handling particularly tricky or ambiguous situations. This ensures that while routine matters might be highly automated, humans are invited to use higher-order skills like critical thinking and creative social communication when the situation warrants.

Worker Inequality

Not coincidentally, the same kinds of skills and capabilities that produce greater worker satisfaction in their exercise— empathy and social communication, critical thinking, problem understanding and response—are precisely those being identified as the last vestiges of the human with respect to automation (Manyika et al. 2017). These higher-order, complex, integrative and deeply human skills—unlike, say, picture annotating or other micro-tasking jobs which are *designed* to be automated as soon as possible—are more likely to be safe from automation far into the future.

This observation directly connects our first concern with alienation to our second concern with economic inequality. The 'Condition of the Turking Class' (Horton 2011) is not simply an experiential problem that describes a particular kind of mindless, repetitive labor. It is also an economic problem, since these kinds of jobs tend to pay very poorly, and are literally just about to be automated. A recent International Labor Organization survey of working conditions covering 3,500 workers living in 75 countries around the world, and working on five English-speaking microtask platforms, found that on average a worker in 2017 earned US$4.43 per hour when only paid work was considered, and US$3.31 per hour when total paid and unpaid hours were considered (Berg et al. 2018). Median earnings were lower, at just US$2.16 per hour when paid and unpaid work were considered.

Conversely, job security from automation is increasingly pegged not just to jobs that are more cognitively difficult, but also jobs where there is variety and integrated functioning. Indeed, since 1980 employment and wage growth has been strongest in jobs that require high levels of both cognitive skill and social skill—again, the variety that makes jobs satisfying, expressing more human skill and, in combination, seeing a greater reward in the marketplace (Deming 2017). The 2019 report of MIT's Work of the Future Task Force echoes these findings, suggesting that policymakers focus on job *quality* rather than job quantity alone, and arguing that countries should concentrate their investments on delivering "middle-skill jobs with favorable earnings and employment security to the vast majority of their workers" (Autor et al. 2019, 17-19).

There is a social cost to making too many jobs that are too elite. In their book the *Second Machine Age*, Brynjolfsson and McAfee (2014) argue that that the growth of inequality today can be directly tied the growth of the tech economy. And just as much as the elimination of routine jobs via automation, the biggest factor in the growing chasm, they argue, is overvaluation of the technology makers: the small elite that innovate and create. The technology-driven economy "favors a small group of successful individuals by amplifying their talent and luck, and dramatically increasing their rewards," (Rotman 2014). We see the results of this in tech-driven economies like Silicon Valley where we work, and where salaries of the class of technical creators are notoriously high and competition for labor is tight, but where other laborers struggle to get by.

The focus on achieving balance between too much and too little autonomy in HAT points us toward the middle: not producing dead-end jobs, but also taking care not to make every human-in-the-loop job into an *engineering* position. If not integrated into other, more lasting tasks in the design of work role, turning the mobility manager exclusively into the role of machine trainer can lead to the problem of temporary, dead-end labor. Conversely, making the human into the role of "technical band-aid" has the potential to eek ever closer to remote engineering. This position implies a relationship to the machine of creation, design, maintenance, or repair, and requires years of specialized training and experience that are out of reach for everyday laborers.

Worker Culpability

Finally there is the issue of liability and blame when something (inevitably) goes wrong. This issue is obviously more serious for safety-critical operations like mobility systems that transport human bodies at high speeds. Self-driving cars are likely to be one of the first intelligent and semiautonomous technologies to be widely adopted in safety-critical

environments. We have yet to see all the ways in which liability will, or will not, be distributed, but we already know that it will be contentious. Culpability is an obvious problem in a legal actor role, but could be an issue in any role for something like an obstructions use case, where operator agency is inserted into decision-making loops that involve live, on-road AV operations. Here again the HAT focus on situational awareness, as well as on transparency and bi-directional teaming, might help.

In a recent paper in *Data & Society*, Elish (2019) describes that intelligent and autonomous systems in every form have the potential to generate "moral crumple zones." A "moral crumple zone" describes how responsibility for an automation error may be incorrectly displaced onto a human actor within the system who in fact had very little control over the erroneous behavior:

> "Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become the component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions. While the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system at the expense of the nearest human operator." (Elish 2019, 40)

The concept of the moral crumple zone ties together the structural and functional features of a system: that is, the complex and unclear distribution of control among multiple actors across space and time, and the popular media's human-centered portrayal of accidents. It explains how human operators come to be primary seats of public accountability in human-machine systems. Moral crumple zones, according to Elish, are likely to take shape in the immediate aftermath of a highly publicized event or accident. And they are also more likely to take place when there are certain disjunctions in the automation paradigm: when there is a mismatch between the capacity of the human-in-the-loop to know about the state of a situation, and the human's authority and capacity to act on that situation.

There are infinite permutations of this disjuncture between acting efficaciously and achieving situational awareness—that is, knowing comprehensively and correctly what is happening and what it means for the future of the system (Endsley 1995). They have played a part in headlining disasters where humans have been dragged through the mud in the media aftermath, including the classic case of the nuclear meltdown at Three Mile Island, as well as the more recent 2018 crash involving an Uber AV in Tempe, AZ, in which a pedestrian was killed. In the latter case, the 'self-driving' car was a modified Volvo XC90 SUV equipped with many driver assistance features, but running Uber's own self-driving software which had (for unclear reasons) disabled those features (NTSB 2018). Had these systems not been disabled, it is expected that the Volvo would have engaged the brakes and stopped before hitting the pedestrian. Yet the report and subsequent media coverage focused on the safety driver's behavior, with concerns raised as to whether the she was looking at her cell phone or streaming media (Somerville & Shepardson 2018). In other words, despite a complex set of factors precipitating the crash, public scrutiny focused on the driver, who may now be facing criminal charges (Elish 2019).

Both safety drivers in autonomous test vehicles and managers at nuclear reactors share a position with respect to automated systems known as "supervisory control." In this paradigm, the autonomous capabilities of the system operate effectively on their own most

of the time, but the system is designed to "hand off" control to the human in the most difficult situations (Sheridan 1992). This might happen when the system recognizes its own fallibility in relation to a difficult situation—such as a nuclear reactor alerting control room operators that something is amiss—or when the human is charged with recognizing an impending issue and 'overriding' the automatic functioning of a system on their own—as is expected of safety drivers in AV systems.

In both cases operators are expected to be alert and monitoring the system, despite few technological affordances supporting the maintenance of that level of mental engagement. The problem with supervisory control, then, relates to one of the "ironies of automation" (Bainbridge 1983) or what Endsley (2017) has called the "automation paradox": as more autonomy is added to a system, and as its reliability and robustness increase, the situational awareness of human operators becomes lower, and it is less likely that they will be able to take over manual control when needed. If the operator is superfluous much of the time, just sitting there watching, this makes it *essentially impossible* to maintain situational awareness. Yet as the "supervisor," the human is in position to be immediately made responsible if they don't 'snap to' and handle those dangerous edge cases appropriately, or proactively detect problems in the automation.

Ultimately, protecting the operator from blame in failure situations will require much more than having the right automation paradigm in place. There must, at minimum, be a policy that accidents are never the human's fault outside of a short list of absolutely essential job requirements, and within the context of specific and known protocols for what the human responsibility is. But given that our intervention in this paper is at the level of the automation paradigm, we can add the requirement that the operator be presented with data consistent with the achievement of situational awareness, and that the work be designed such that their 'human factors' are respected enough to keep them engaged to a degree commensurate with their moral and legal responsibility.[2] In other words, what is most important is not that the human have "more" agency in situation so they can "take the wheel" when needed. Rather, what matters in work design for highly automated systems is that there is congruence between awareness and responsibility, and enough transparency for the operator to understand what the automation is doing and what she can do to affect it.

HAT: A MORE ETHICAL AUTOMATION PARADIGM?

Taking these three issues—alienation, inequality, and culpability—together, we get a picture of a position that we would like to design that can be described in terms of a few organizing values. This is a position characterized by variety of tasks, continuous engagement in knowledge-gathering and decision-making, and congruence between awareness and responsibility. Rather than focusing on making the Mobility Manager a social-knower, legal entity, or machine trainer, the best outcome for the worker might be to have them engage in all of these different roles at different moments in a work flow, and to play these roles at different levels of automation *vis-à-vis* the machine, and in different ways. Variety, in particular, would seem to emerge as a clear winning value: it makes the job less liable to be automated in the future, and thus potentially higher skilled and more humane; and it might also engage the worker more, keeping her cognizant of her level of responsibility and perhaps more interested in the task.

Our argument is that there is a potential congruence between HAT principles for creating operational effectiveness through an intermediate-automation approach—where an operator is working on a variety of kinds of situations, and at a variety of levels of automation, while maintaining situational awareness—and worker well-being on the job in a more wholistic sense. Although it is in its relative infancy as an automation paradigm, HAT seems to be a more humane and plausible vision than other automation paradigms being pursued, within and without our organization. By operating under the rubric of Teaming, we have been able to make technical and safety arguments for certain relations to the machine that we consider potentially more ethical, and which might result in a job that is engaging, at a medium skill level, and that could protect the operator from mismatch between what they know and what they are capable of doing (and from resultant blame for accidents).

Obviously none of this can guarantee a "good job," nor can it shield the operator from blame if something goes wrong absent larger institutional and social protections. Further, HAT is minimally-developed on a technical level, and requires continued research and testing. But our hope is that in continuing to use this paradigm to experiment with the calibration of human agency in effective coordination with AI in our SAM system, in a terrain where the *what* and *how* of the human being's involvement is so up in the air, we can push for a more progressive worker agenda. We are finding in the "team" an ability to focus on technical performance while maintaining (sometimes covert) attention to human well-being.

In our business, the argument must be made that retaining human dignity will make workers more productive in creating business value, or that efficient management of highly automated systems is simply impossible without agential, empowered humans in the loop.[3] Rather than forwarding purely ethical arguments for the higher-order functioning, diversity of tasks, and other desirables that we think are consistent with better overall outcomes for workers, Teaming has provided us with a technical and theoretical basis to argue these are necessary to system operations. Luckily, through collaborations with the open-minded engineers, designers and project managers with whom we have the privilege of working on Mobility Management, this Teaming vision seems to be winning for now over other contending automation paradigms at our lab.

Laura Cesafsky is a UX Researcher at the Alliance Innovation Laboratory in Silicon Valley. Their work focuses on shaping the user experience of workers, customers and publics as they interact with AI-intensive vehicle systems.

Erik Stayton is a PhD candidate in the Program in History, Anthropology, and Science, Technology and Society at MIT. He investigates human interactions with AI systems, and currently studies the values implicated in the design, regulation, and use of automated vehicle systems. He has been an intern at the Alliance Innovation Laboratory.

Melissa Cefkin is Principal Researcher and Senior Manager of the User Experience group at the Alliance Innovation Laboratory in Silicon Valley. She has had a long career as an anthropologist in industry, including time at the Institute for Research on Learning, Sapient, and IBM Research.

NOTES

Acknowledgments – We thank our colleagues at the Alliance Innovation Lab in Santa Clara, CA—especially the Seamless Autonomous Mobility (SAM) team—for their support and collaboration in this research. We also appreciate the detailed commentary and guidance from our EPIC reviewers, both named and anonymous, in the shaping of this article.

1. The Seamless Autonomous Mobility system was first publicly demonstrated at the Consumer Electronics Show in 2017. Videos and press images of the system are available online.

2. Designing for operator engagement—up to and including feeding unnecessary or non-critical tasks to keep the operator aware—is an important part of Joint Cognitive Systems Design, and is used in airline contexts to maintain pilot situational awareness (Woods and Hollnagel 2006).

3. For more on this topic, see our previous EPIC paper focused specifically on what benefits beyond operational capabilities alone that empowered human beings can bring to a system (Stayton and Cefkin 2018).

REFERENCES CITED

Aitken, H. G. J. 1960. *Taylorism at Watertown Arsenal: Scientific Management in Action.* Harvard University Press, Cambridge, MA.

Autor, David, David Mindell, and Elisabeth B. Reynolds. 2019. *The Work of the Future: Shaping Technology and Institutions.* Massachusetts Institute of Technology. https://workofthefuture.mit.edu/report/work-future

Bainbridge Lisanne. 1983. "Ironies of Automation." *Automatica* 19 (6): 775-779.

Berg, Janine, Marianne Furrer, Ellie Harmon, Uma Rani, and M. Six Silberman. 2018. *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World.* Geneva: International Labour Organization.

Blomberg, Jeanette. 2005. "The Coming of Age of Hybrids: Notes on Ethnographic Praxis." *Ethnographic Praxis in Industry Conference Proceedings*, 67- 74.

Bradshaw, Jeffrey, Robert R. Hoffman, Matthew Johnson, and David D. Woods. 2013. "The Seven Deadly Myths of "Autonomous Systems." *IEEE Intelligent Systems* 28 (3): 54-61.

Brandt, Summer, Joel Lachter, Ricky Russell, and Robert Jay Shively. 2017. "A Human-Autonomy Teaming Approach for a Flight-Following Task." *Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering.*

Brynjolfsson, Erik and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress and Prosperity in a Time of Brilliant Technologies.* New York: W. W. Norton and Company.

Callon, Michel. 1984. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay." *The Sociological Review*, 32: 196-233.

Cefkin, Melissa, Obinna Anya and Robert Moore. 2014. "A Perfect Storm? Reimagining Work in the Era of the End of the Job." *Ethnographic Praxis in Industry Conference Proceedings,* 3–19.

Clancey, William J. 2014. *Working on Mars: Voyages of Scientific Discovery with the Mars Exploration Rovers.* Cambridge, MA: MIT Press.

Davies, Alex. 2018. "Self-Driving Cars Have a Secret Weapon: Remote Control." *Wired* website, January 2. Accessed October 30, 2019. https://www.wired.com/story/phantom-teleops/

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *The Quarterly Journal of Economics*, *132* (4): 1593-1640.

Diebold, John. 1959. *Automation: Its Impact on Business and Labor.* Washington, D.C.: National Planning Association.

Endsley, Mica R. 1995. "Toward a Theory of Situational Awareness in Dynamic Systems." *Human Factors Journal* 37 (1): 32-64.

Endsley, Mica R. 2017. "From Here to Autonomy: Lessons Learned from Human-Automation Research." *Human Factors* 59 (1): 5-27.

Elish, Madeline. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." In *Engaging Science, Technology, and Society* 5: 40-60

Gray, Mary L., & Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* New York: Houghton Mifflin Harcourt.

Harris, Mark. 2018. "Waymo Filings Give New Details on Its Driverless Taxis." *IEEE Spectrum* website, May 14. Accessed October 30, 2019. https://spectrum.ieee.org/cars-that-think/transportation/self-driving/waymo-filings-give-new-details-on-its-driverless-taxis;

Horton, J. J. 2011. "The Condition of the Turking Class: Are Online Employers Fair and Honest?" *Economics Letters*, 111 (1): 10-12.

Irani, Lilly. 2015. "Justice for Data Janitors." *Public Books*, January 15. Accessed October 30, 2019. https://www.publicbooks.org/justice-for-data-janitors/

Kockelman, Paul. 2007. "Agency, The Relation between Meaning, Power and Knowledge." In *Current Anthropology* 48 (3): 375-401.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network Theory.* Oxford, UK: Oxford University Press.

Law Commission. 2018. *Automated Vehicles: A Joint Preliminary Consultation Paper*. Consultation Paper 240. (Scottish Law Commission Discussion Paper 166.) November 8. https://www.lawcom.gov.uk/project/automated-vehicles/

Lehdonvirta, Vili. 2016. "Algorithms that Divide and Unite: Delocalisation, Identity and Collective Action in 'Microwork'." In *Space, Place and Global Digital Work*, edited by Jorg Flecker, 53-80. London: Palgrave Macmillan.

Licklider, J.C.R. 1960. "Man-Computer Symbiosis." In *IRE Transactions on Human Factors in Electronics*, HFE-1 (March): 4-11.

Manyika, James, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko and Saurabh Sanghvi. 2017. "Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills, and Wages." *McKinsey Quarterly,* November. Accessed October 30, 2019. https://www.mckinsey.com/featured-insights/future-of-work/jobsvc-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

Margeit, Robert. 2019. "Fully Autonomous Driving Will Still Need the Human Touch – Nissan." *CarAdvice* website, March 17. Accessed October 30, 2019. https://www.caradvice.com.au/735920/nissan-seamless-autonomous-mobility/

Markoff, John. 2014. "Google's Next Phase in Driverless Cars: No Steering Wheel or Brake Pedals." *New York Times* website, May 27. Accessed October 30, 2019. https://www.nytimes.com/2014/05/28/technology/googles-next-phase-in-driverless-cars-no-brakes-or-steering-wheel.html

Marx, Karl. 1844. "Estranged Labor" *Economic and Philosophic Manuscripts.* First published 1932. Quotation from Moscow: Progress Publishers. 1959. https://www.marxists.org/archive/marx/works/1844/manuscripts/preface.htm

Mindell, David. 2011. *Digital Apollo: Human and Machine in Space- flight.* Cambridge, MA: MIT Press.

Mindell, David. 2015. *Our Robots, Ourselves: Robotics and the Myths of Autonomy.* New York: Viking.

Nissan Motor Corporation. 2017a. Seamless Autonomous Mobility (SAM) webpage. Accessed October 30, 2019. https://www.nissan-global.com/EN/TECHNOLOGY/OVERVIEW/sam.html

Nissan Motor Corporation. 2017b. "Seamless Autonomous Mobility: The Ultimate Nissan Intelligent Integration." *Nissan Channel 23 Blog,* January 8. Accessed October 30, 2019. https://blog.nissan-global.com/EN/?p=14313

Noble, David F. 1984. *Forces of Production: A Social History of Industrial Automation.* New York, NY: Knopf.

NTSB. 2018. *Preliminary Report Highway: HWY18MH010.* National Transportation Safety Board report. https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18MH010-prelim.aspx

Schaefer, Kristin E., Edward R. Straub, Jessie YC Chen, Joe Putney, and Arthur W. Evans III. 2017. "Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams." *Cognitive Systems Research* 46: 26-39.

Seeber, Isabella, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, and Matthias Söllner. 2019. "Machines as Teammates: A Research Agenda on AI in Team Collaboration." In *Information and Management*. https://doi.org/10.1016/j.im.2019.103174

Sheridan, Thomas. 1992. *Telerobotics, Automation, and Human Supervisory Control.* Second edition. Cambridge, MA: MIT Press.

Shively, Robert Jay, Joel Lachter, Summer L. Brandt, Michael Matessa, Vernol Battiste, and Walter W. Johnson (2017). "Why Human-Autonomy Teaming?" *Proceedings of International Conference on Applied Human Factors and Ergonomics*: 3-11.

Simondon, Gilbert. 2011. "On the Mode of Existence of Technical Objects." *Deleuze Studies*, *5* (3): 407-424.

Smith, Merritt Roe. 2016. "Yankee Armorers and the Union War Machine." In *Astride Two Ages: Technology and the Civil War*, edited by Barton Hacker, 25-54. Smithsonian Institution Press.

Somerville, Heather and David Shepardson. 2018. "Uber Car's 'Safety' Driver Streamed TV Show Before Fatal Crash: *Police*." *Reuters*, June 21. Accessed October 30, 2019. https://www.reuters.com/article/us-uber-selfdriving-crash-idUSKBN1JI0LB

Stayton, Erik and Melissa Cefkin. 2018. "Designed for Care: Systems of Care and Accountability in the Work of Mobility." *Ethnographic Praxis in Industry Conference Proceedings:* 334-350.

Taylor, Frederick Winslow. 1911. *The Principles of Scientific Management*. New York and London: Harper & Brothers.

White, Lynn, Jr. 1962. *Medieval Technology & Social Change* London, England: Oxford University Press.

Woods, D. D. and Hollnagel, E. 2006. *Joint Cognitive Systems: Patters in Cognitive Systems Engineering.* CRC Press, Taylor & Francis, Boca Raton, FL.

Yapchaian, M. "Human-Centered Data Science: A New Paradigm for Industrial IoT." *Ethnographic Praxis in Industry Conference Proceedings*: 53-61.