UCLA
AI PULSE Papers

Title
AI & Agency

Permalink
https://escholarship.org/uc/item/8q15786s

Authors
Newman, Sarah
Birhane, Abeda
Zajko, Mike
et al.

Publication Date
2019-09-26

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

Introduction

In July of 2019, at the Summer Institute on AI and Society in Edmonton, Canada (co-sponsored by CIFAR and the AI Pulse Project of UCLA Law), scholars from across disciplines came together in an intensive workshop. For the second half of the workshop, the cohort split into smaller working groups to delve into specific topics related to AI and Society.

I proposed deeper exploration on the topic of "agency," which is defined differently across domains and cultures, and relates to many of the topics of discussion in AI ethics, including responsibility and accountability. It is also the subject of an ongoing art and research project I'm producing. As a group, we looked at definitions of agency across fields, found paradoxes and incongruities, shared our own questions, and produced a visual map of the conceptual space. We decided that our disparate perspectives were better articulated through a collection of short written pieces, presented as a set, rather than a singular essay on the topic. The outputs of this work are shared here.

This set of essays, many of which are framed as provocations, suggests that there remain many open questions, and inconsistent assumptions on the topic. Many of the writings include more questions than answers, encouraging readers to revisit their own beliefs about agency. As we further develop AI systems, and refer to humans and non-humans as "agents"– we will benefit from a better understanding of what we mean when we call something an "agent" or claim that an action involves "agency." This work is under development and many of us will continue to explore this in our ongoing AI work.

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

– Sarah Newman, Project Lead, August 2019

1. Characterizing Agency

Jon Bowen
PhD student in Philosophy, Western University

Some of the beings we encounter in our environment are inanimate. These things may be pushed and pulled, they may collapse or disintegrate. In each of these cases, the entities are fundamentally passive–if they move or change, one suspects that these movements and changes will be exhaustively explained by appealing to mechanical forces within or without.

But there is another kind of entity in our environment. These beings seem to be fundamentally goal-directed. To appearances, they are spontaneous initiators of their own actions. These are animate beings, or agents. The movements of these entities seem to be best explained not by appeal to mechanical causes of their activity, but to the goals that they are striving towards, the beliefs they have about the world, and their desires.

Giving a precise definition of what animacy or agency consists of is no easy task for the philosopher, but nonetheless we appear to have no difficulty at all recognizing animate motion and distinguishing it from the motion of inanimate objects. Even human infants, it seems, can detect animate motion and differentiate it from inanimate motion in point-light displays, even when occlusions are present.

But why should this be the case? Why would it be so difficult to give a theory of intentional action, and yet so easy to detect it? I will set out one suggestion. We do not, as has been proposed, infer intentions, beliefs, and desires as a part of a theory

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

for explaining or predicting behavior. Instead, intentional action is behavior with certain distinctive, overt characteristics, which our perceptual systems have evolved to directly perceive. Goal-directed behaviors, I will suggest, are a very real kind of behavior out there in the world with distinctive characteristics. Furthermore, it is important that animals perceive and understand this particular kind of behavior, and sure enough, they are able to do so with astonishing acuity.

What are we saying when we explain the activities of another person (or of a non-human animal) by appealing to their intentions? Here I will draw on an analysis from Dennis Walsh: "A teleological explanation is one that explains the nature or activities of an entity, or the occurrence of an event, by citing the goal it subserves. A system has goal, E, just in case it exhibits goal-directed behavior toward E. Goal-directed behavior is a gross property of a system as a whole." (p. 177)

What this amounts to is not an account of the intrinsic causal etiology of the agent's behavior. Instead, we are locating that behavior in a chain of events that show a certain distinctive pattern. If an agent is trying to do X, then its behavior will flexibly reconfigure itself in the service of that goal. When a dropped object encounters the ground, it will stop. When an agent's initial attempts to pursue some goal are thwarted, that agent will spontaneously and flexibly reconfigure its behavior so as to continue to pursue its goal. A human need not stop at the ground–they can retrieve a shovel, and perhaps a jackhammer or a drill if called for (if they really want to!) This is to say, when an agent is engaging in goal-directed activity, its behavior is robust against perturbations and obstacles in a way characteristically not present in inanimate objects.

If there are such systems in nature–systems that will reliably produce effects by marshaling their intrinsic causal capacities in the service of goals–then clearly the

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

perceiving animal would be at an advantage if they could detect them when present! The challenge, from the perspective of an animal's perceptual system, then, is to detect or pick up the information which specifies what the goals of other agents in their acting are. While this might sound like quite a feat, again, this is something we all seem to be very good at.

If agency amounts to the capacity for intentional action, and the preceding account of goal-directed behavior is sound, what basis might there be to deny that such a thing exists as a real phenomenon in nature, and a real attribute of natural beings?

2. The Value of the Concept of Agency in an Increasingly Rational World

Osonde Osoba
Information scientist, RAND Corporation
Professor, Pardee RAND Graduate School

Let us concede that different traditions of thought have different definitions and perspectives on what it means to be an agent or to have agency. There are some common threads that may be useful to highlight. I will focus on one. Most conceptions of Agency are rooted in action, in doing, in affecting a substrate environment.

A working definition for the purposes of this discussion could go thus:

An agent is an entity that is capable of causing or effecting change in its world in pursuit of private (personal) goals.

This definition has a couple of features worth highlighting:

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

The primacy of causality: We focus on the idea of causal influence as a defining characteristic. An entity whose whole existence consists of internal ruminations (e.g. Ibn Tufayl's floating man) does not meet our criteria. However much sophisticated intelligence it applies to its sense perceptions, it has no influence over its environment. It can achieve no goals in its world no matter how intensely it wills them.

Contextual worlds: Context determines the relevant world over which the agent aims to exert influence. Entities can be part of numerous worlds or environments. An entity's agency in each of these worlds is determined by how much causal influence it can exert in each one. We can imagine a measure of power based on what fraction of an agent's environment it can influence.

Private goals: Private goals may be related to Aristotle's idea of a "final cause," the reason for which a thing exists. The capacity for pure action without goals requires no planning, interiority, or intentionality. We will argue that tracking that sort of capacity is not useful.

The concept of agency has proven useful for rooting responsibility and/or liability in entities capable of modifying their actions in response to external influence. Such a capacity for redress or accountability can arguably only be supported by entities capable of goal-oriented behavior.[1] Responsibility can be moral or legal (more coercive/backed by institutional power). Agency likely serves other important functions. But the responsibility-rooting function of agency is crucial for influencing or controlling behavior in social structures.

This view of agency is explicitly not about independence or autonomy. Agency, in this conception, is closer to a useful fiction that enables the clean assignment of

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

responsibility and dessert. And the default assumption is that agents exist within networks of influence. A degree of external manipulation of agents is the norm, not a novel pattern.

Historically, the use of agency for allocating moral responsibility has been a useful but imperfect device: the assignment of moral responsibility has not always tracked causal responsibility. The long tradition of arguments for the justice of gods (theodicies) is a case in point. If evil befalls a person, it must be because that person has misused his agency ("sinned") and therefore deserves or is morally responsible for his lot.[2] Some superstitions may also be construed to serve a similar function. These failures in causal attribution happen because the world is complex, causal attribution is notoriously difficult, & causal influences can be very subtle when they exist. By contrast, gods are simpler, more convenient causal explanations.

Our modern conception of moral responsibility is becoming more rational, more scientific. Part of the goal of rational thought is to focus on the true causes of observed phenomena. Weber goes so far as to argue that scientific inquiry is just a rational incarnation of theodicy.[3] We have moved from agency based on imperfect beliefs towards a more causal conception of responsibility.

But what happens when our rational understanding of reality expands to the point where we are able to track causal influences as finely as possible?[4] E.g. recent literature has begun to undermine agency-based explanations of individual behavior in favor of longer chains of causal influence that reach past the mask of more person-focused conceptions of agency. How do we ground responsibility and liability when large swathes of action can be explained away via causal factors outside the individual (e.g. the larger explaining value of social influence or manipulation,

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

genetics, environmental factors, etc.)?

Does the concept of agency survive this trend?

3. Human agency in the age of AI

Abeba Birhane
PhD Candidate, School of Computer Science, University College Dublin

Provocation:

The question of agency necessarily provokes the question of what it means to be a person and, in particular, what it means to be a person in the age of ubiquitous artificial intelligence (AI) systems. We are embodied beings that inherently exist in a web of relations within political, historical, cultural, and social norms. Increasingly, seemingly invisible AI systems permeate most spheres of life, mediating and structuring our communications, interactions, relations, and ways of being. Since we do not exist in a social, political, historical, and AI-mediated vacuum, it is imperative to ground agency as inherently inseparable from the person as construed in such contingent constituent factors. Depending on the context and the space we occupy in the social world, all these dynamic and contingent factors serve as enabling constraints for our capacity to act. Our capacity to act within these contextual factors varies in degree depending on the space we occupy at a certain time, in a certain socioeconomic context; the more privileged we are, the fewer the potential constraints, and the greater our degrees of agency.

Essay:

The individual is never a fully autonomous entity: rather, they come into being and maintain that sense of existence through dynamic, intersubjective, and reciprocal

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

relations with others.[5] Our biology, current social and cultural norms, historical, and contextual contingencies, as well as our physical and technological environment, constitute who we are and our degrees of agency within a given time and context. Increasingly, AI systems are becoming an integral part of our environment – be it the search engines that we interact with, our social media activities, the facial recognition systems that we come in contact with, or the algorithmic systems that sift through our job applications – further adding enabling, or limiting, constraints. (Enabling constraints here might include having a common Western male name, or other demographic traits, that the job application algorithm chooses to include, rather than exclude. These are still constraints, but in certain instances they increase opportunity, rather than decrease them.)

We are embodied beings that necessarily exist in a web of relations with others, within certain social and cultural norms as well as through emerging technologies. This means our sense of being, as well as our capacity to act, are inextricably intertwined and continually changing as we move between various spheres taking on various roles. The various factors that constitute (and sustain) who we are influence the varying degrees of agency we are afforded. As we go on about our daily lives, we move between various social and cultural conventions, physical environmental enablers (or disablers) of certain behaviors and actions (as opposed to others), and technological tools that shape, reinforce, and nudge behavior and actions in certain directions (and not others). As a PhD student, my role, responsibility, and capacity to act in my academic environment, for example, is different than that of my role, responsibility, and capacity for action when I am at a social gathering within the immigrant community. Furthermore, my interaction with others through Twitter is different from both these other contexts, and is partially determined by the ways the medium affords possible actions and interactions. Our sense of agency, then, is fluid,

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

dynamic, and continually negotiated within these various physical, mental, psychological, technological, and cultural spaces. Discussion of agency, consequently, cannot emerge in a social, technological, and contextual vacuum. Nor is it something we can view as stable or pin on individual persons due to the complex, contingent, and changing factors that constitute and sustain personhood.

Conversely, agency cannot be an abstract term that we attempt to define and analyze in a general, one-size-fits-all manner but one that needs to be grounded in people. People, due to their embeddedness in context, culture, history, and socio-economic status, are afforded varying degrees of enabling constraints. Agency, therefore, is not an all-or-nothing phenomenon but something that varies in degrees depending on individual factors, circumstances and situations. Individuals at the top of the socio-economic hierarchy, for example, face relatively fewer disabling constraints, consequently resulting in a higher degree of agency, and the reverse holds for those at the lower end of society. For example, depending on their socio-economic and educational background, one may be labelled “eccentric” vs. “insane”, a “lone wolf” vs. a “radicalized extremist”, a “freedom fighter” vs. a “terrorist”.

Agency, AI, and ethical considerations

Living in a world of ubiquitous networked communication, a world where AI technologies are interwoven into the social, political, and economic sphere means living in a world where who we are, and subsequently our degree of agency, is partially influenced by automated AI systems.

The concept of AI often provokes the idea of (future and imaginary) sentient artificial beings, or autonomous vehicles such as self-driving cars or robots. These preconceptions often assume (implicitly or otherwise) that AI systems are entities that exist independently of humans in a machine vs. human dichotomy. This view,

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

which dominates academic and public discourse surrounding AI is a deeply misconceived, narrow, and one-dimensional conception of AI. What AI refers to in the present context is rather grounded in current systems and tools that operate in most spheres of life. These are seemingly invisible tools and systems that mediate communication, interaction with others and other technological infrastructures that alter the social fabric. These AI systems make life effortless, as they disappear into the background to the extent that we forget their very existence. They have become so inextricably integrated with our daily lives that life without them seems unimaginable. As Weiser[6] has argued, these are the most profound and powerful technologies. "The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it."

These systems sort, classify, analyze, and predict our behaviors and actions. Our computers, credit card transactions, phones, and the cameras and sensors that proliferate public and private spaces are recording and codifying our "habits", "behaviors", and "experiences". Such ubiquitous interlinked technological milieu continually maps out the where, when, what, and how of our behaviors and actions, which provide superficial patterns that infer who we are.[7] Whether we are engaging in political debate on Facebook, connecting to "free" wi-fi, using Google Maps to get from point A to B, searching for sensitive health information on Google, ordering grocery shopping, posting selfies on Instagram, or out in the park for a jog; our actions and behaviors produce a mass flow of data that produce pattern-based actionable indices about "who we are". These superficial extrapolations, in turn, feed models that predict how we might behave in various scenarios, whether we might be a "suitable" candidate for a job, or are likely to commit crimes, or are risks that should be denied loans or mortgages. Questions of morality (often misconceived as

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

technical questions in need of a technical fix) are increasingly handed over to engineers and commercial industries developing and deploying AI systems as they are bestowed with sorting, pattern detecting, and predicting behaviors and actions. These predictive systems give options and opportunities to act or they limit what we see and the possible actions we can take. And as O'Neil[8] reminds us, each individual person does not pass through these processes to the same degree nor do they suffer the consequences equally. "The privileged are processed by people, the masses by machines."

These systems not only predict behavior based on observed similar patterns, they also alter the social fabric and reconfigure the nature of reality in the process. Through "personalized" ads and recommender systems, for example, the level and amount of options put in front of us varies depending on the AI's decision of "who we are," which reflects the place we occupy in the social hierarchy. The constraints that provide us with little or great room to act in the world are closely related to our socio-economic status and, increasingly, to who our data says we are. Unsurprisingly, the more privileged we are, the more we are afforded the capacity to overrule algorithmic identification and personalization (or not be subjected to them at all), maximizing our degrees of agency.

Since agency is inextricably linked to subjecthood, which is necessarily political, moral, social, and increasingly digital, the impact of power structures is inescapable. These power relations and the capacity to minimize the potential constraints AI imposes on agency, is starkly clear when we look at the lifestyle choices that powerful agents in Silicon Valley, who make and deploy technology, are afforded. For example, while screen-aided education is pushed towards mainstream schools, the rich on the other hand are reluctant to adopt such practices.[9] Agency, the

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

capacity to act in a given technological environment and context varies in degree from person to person. Silicon Valley tech developers, those with power and awareness of technology as constraining powers, are reluctant to let it infiltrate their children's surroundings. Some go so far as banning their nannies from the use of screens.[10]

Agency is not an all-or-nothing phenomenon that we either do or do not have. Rather, agency is inextricably linked to our social, political, and historical contexts, which are increasingly influenced by technological forces. These forces grant people varying degrees of agency. In an increasingly AI-powered society our capacity to act is limited or expanded based on our privilege; agency is increasingly becoming a commodity that only the privileged can afford.

4. Agency to Change the World

Mike Zajko
Assistant Professor, Department of History and Sociology, University of British Columbia

Abstract
Social theory has identified agency with social change and dynamism, bringing tension and possibility to a world where social structures are reproduced. The concept of agency can rescue us from the notion that we are simply the product of our conditioning (zombies of embodied habits), and stands in opposition to ontologies that foreground practices at the expense of subjects. While a humanist conception links agency to purposive action, an expansive (post-humanist) definition elides the question of intentionality, and links agency with action, irrespective of purpose. According to this view, rather than being an exclusive human property,

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

agency is all around us, and society has always consisted of relations between human and non-human actors. We should keep in mind that agency is not absolute or independent, but contextual and relational. If we conceive of agency in this way, we can see the stakes of some of the current debates about AI: to what extent will these systems act as agents of change in our world, and how will AI affect (enhance, extend, supplant, or constrain) human agency?

Agency to Change the World

How did Western intellectuals go from believing agency to be the exclusive property of the human subject, to considering whether algorithmic agents, or AI systems, also have agency? One understanding is that as AI increasingly approximates human intelligence, it attains attributes formerly reserved for humanity. But an argument can be made that AI today, even in its narrowest forms, already exercises agency, and that humans were never particularly special to begin with.

It is commonly said that people exercise agency to achieve their desires, goals, and interests. In sociology, agency has long been seen as the source of change in society. Agency is why society does not remain in a steady state, despite all the ways that social structures are reproduced. Without agency, we would all be pawns shaped and manipulated by larger forces that often precede our existence: children molded into reproductions of their parents; compliant, orderly workers reproduced by the educational system to passively accept ideologies that justify why the existing order is natural, desirable, or worthy of being preserved. Agency refers to our ability to change this social structure, to disagree with our parents, use education to advance knowledge, achieve social mobility, critique ideology, and challenge government.

There is a longstanding debate in social theory about the relationship between agency and social structure,[11] which has largely gone stale and unresolved. But

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

agency continues to provide the tension that prevents a totalizing view of structure. Not everyone agrees that agency is required to understand humanity or the relationship between individuals and society, but social theories that do without agency, or that provide an impoverished view of agency, paint a deterministic picture. Individuals are conceived not as subjects, but through their habits and practices, or as the effects of the social structures that produce them. Without agency, we are zombies, automata, or cultural dupes.[12]

Amidst some of the debates about agency and structure in the 1980s and 1990s, a new conception (often associated with Bruno Latour and Actor-Network Theory)[13] began to take hold. The provocative argument was that agency was not confined to humans, but that society was composed of both human and non-human "actants."[14] Agency was defined roughly with action, and the ability to affect the world. If a human worker was replaced by an object (even an inanimate one)[15] that could play the same role, then that object similarly exercised agency. Because in many of our interactions with the world, whether in laboratory experiments or farming,[16] humans cannot fully predict or control the outcome, nature also has agency – co-creating the world with us. Questions of intentionality and purposiveness are elided through this focus on action.

Whether in its humanist or post-humanist form, Western theory's interest in agency has also been subjected to significant critique. The idea of an autonomous human subject is arguably a historical invention – a distinctly Western, masculine, individualistic vision of man. Feminist theorists advanced these arguments decades ago, pointing to the often unacknowledged work (disproportionately performed by women) of nurturing and caring for 'autonomous' subjects. Complicating but not

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

necessarily rejecting the ideal of autonomy, these authors advanced a concept of agency that situates it firmly within social relationships.[17] Relational and non-human conceptions of agency are common in Indigenous and non-Western ontologies,[18] rooted in the understanding that the world is agentially alive, and that humanity is inexorably linked to and dependent on these forces. In an article section titled *Columbus Discovers Non-Human Agency*,[19] three authors influenced by Indigenous feminist literature point to the Eurocentric and settler colonial bias of a recent turn in social theory. In this 'new materialism', authors influenced by Latour and feminist STS have made expansive claims about agency that may be innovative for social theory, but which are quite traditional for unacknowledged indigenous ontologies.

At this point it is worth reflecting on these divergent conceptions of agency. Along one dimension outlined above, they run the range from treating agency as a distinctly human property, linked to subjectivity, consciousness and intentionality, to a broader view of agency as whatever has effects on the world (and ourselves). At its broadest, we are not agents at all: distinctions of subjects and objects are dissolved, and the entire universe becomes a quantum soup of intra-active becoming.[20] But somewhere between this posthuman extreme and the reassurance of conventional humanism, we can return to a view of agency that encompasses both humans and AI, as agents that change the world, and are entwined in relations with one another.

Today, developers are building robots that learn about and interact with their environment – an environment that includes other robots as well as humans. Machine learning enables AI systems to pursue goals in ways that humans could not anticipate, even if their goals were initially formulated by humans. We now regularly interact with various kinds of AI, or are subject to decisions made by these systems. Finally, the distinctiveness or exceptionality of the human subject has been

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

repeatedly problematized by advances in AI and in our understanding of other organisms. In this context, conceiving of agency as the ability to change the world remains valuable for considering issues common to humanity and AI.

Conceptualized as a means of social change, we can see that agency is not a human birthright, and is not equally distributed across humanity. Structured inequality provides opportunities to some, which are denied to others. Where a person is born, and how they are nurtured or socialized, has great consequences for the choices and capacities available to them – including the impact a person can have on reshaping pre-existing structures. Agency depends on our relationship to these structures, as well as to each other. Hence, agency varies across positions in society and is subject to change. We can engineer technologies and social systems to enhance human agency, to provide capabilities for transformation of individual or collective conditions; or we can design to preserve and reinforce existing power structures. Similarly, it is valuable to conceptualize the agency of AI through its ability to affect the world, change itself, and change human lives, irrespective of consciousness or intentionality. If we conceive of agency in this way, we can see the stakes of some of the current debates about AI: to what extent will these systems be agents of change in our world, and how will AI affect human agency? What decisions will AI make on behalf of humans, and how will these sociotechnical systems reconfigure the possibilities available to us?

5. Can (and Should) AI Be Considered an Agent

Gabriel Lima
Computer science undergraduate student, KAIST, South Korea

Provocation

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

In this short essay, I share my thoughts on the relationship between artificial intelligence (AI) and various definitions of agency. Can AI be considered an agent? More specifically, does AI fulfill requirements set forth in various definitions of agency? Depending on the perspective and definition taken by the reader, the agency of AI could be controversial, unimaginable, or an unquestionable truth. A question that is often neglected, however, is whether AI should be given any agency. Even though we often derive normative statements of value (e.g., should, ought to) from descriptive statements of fact (e.g., can, is), their distinction is extremely important and has been discussed by many philosophers who argue this relation is not necessarily valid and advisable. Finally, I conclude my essay by raising the open question whether AI should indeed be an agent in our society independent of the fulfilment of agency requirements set by various definitions. Instead of focusing on the abilities of an AI, what if we first ask whether it would be beneficial to treat an AI as an agent in society?

Introduction

Agency has never been clearly defined across, or even within, disciplines. Even though it is often related to autonomy, responsibility, or causality, no clear definition agrees on every detail around the complicated issue of who (or what) is an agent.

In this short essay, I share my thoughts on the relationship between artificial intelligence (AI) and agency. Can AI be considered an agent? More specifically, does AI fulfill requirements set in various definitions of agency? Depending on the perspective taken by the reader, the agency of AI could be controversial, unimaginable, or an unquestionable truth. A question that is often neglected, however, is whether AI should be given agency. Even though we often derive normative statements of value (e.g., should, ought to) from descriptive statements of fact (e.g., can, is), their distinction is important and many philosophers have argued

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

that this connection is not valid or advisable. Finally, I conclude my essay by raising the open question whether AI should indeed be an agent in our society independent of the fulfilment of agency requirements set by various definitions.

As introduced above, agency is not clearly defined and thus, tackling whether AI could qualify as an agent following every single proposed idea of agency is infeasible. In the following short subsections, I will deal with some common sociological, legal, philosophical, and technological definitions of agency and share my thoughts on whether AI could be considered an agent under each definition.

An Agent Is a Goal-Oriented Entity

Does an AI have a goal? From a computer science perspective, this is often how we create and train AIs. For instance, in reinforcement learning we teach AIs by rewarding them depending on whether or not they have achieved a set goal. The goals of an AI are not intrinsic, but extrinsic; the programmer sets its goals following his or her needs. This does not, however, disqualify AI as a candidate for agency. According to the idea that agency is based on a goal-oriented behavior, AI could be seen as an agent.

An Agent Can Act and Modify Its Behavior Depending on the Environment

This definition is often used in computer science when dealing with reinforcement learning, a method used to train AIs. In this setting, we define AI as an agent in an environment with a set of policies and actions. Given that AI is defined as an agent from its conception, it is easy to imagine an AI as an agent after its deployment.

An Agent Has an Effect on the World and Drives Social Change

Following this more sociological perspective, an agent must make a difference in society to qualify for agency. In the current "AI Summer," AI is affecting society in ways many did not expect – or did expect, but unfortunately neglected. AI has been

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

disruptive in diverse sections of society. Job markets having to adapt to the insertion of these electronic entities, and recommendation algorithms controlling what kind of information a certain part of society has access to, are among many examples of novel consequences AI is imposing on society. It is not hard to see an AI as an agent considering its impact on society.

An Agent Can Engage With or Resist Colonial Power

Even though sci-fi scenarios give us the idea that AI can resist the power of its creators, this possibility is far from us. AI cannot resist and turn against its own creator, due to both lack of ability and the high level of control creators still have over their creations. AIs are distant from engaging with (or inverting) the power pyramid, where they are at the very bottom. More importantly, how can they even set that as a goal, if an AI is not currently able to have intrinsic goals? By this conception, AI cannot be an agent since it does not engage in any action dealing with its creators and its hierarchical position.

An Agent Is an Entity That Acts on Behalf of a Principal

We often build AIs as entities to complete a certain task for humans. These systems act on behalf of a principal, which can be their programmers, manufacturers, or users. The principal sets the AI's goals and the system works towards achieving them. By this conception of agency, an AI is clearly an agent. Some authors even argue that AI could be a "perfect agent," since it does not have intentions or goals that could deviate from its principal's goals.

As issued raised by many legal scholars about AI agency is the usual requirement of a contract to establish a principal-agent relationship. Since AI has not (yet) been granted any kind of legal personhood, it cannot be a party to a legal contract. Consequently, while an AI could be seen as an agent under a principal in economic

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

terms, it cannot qualify as one legally.

An Agent Can Bear Responsibility for Its Actions

Can an AI be responsible for its actions? How would this responsibility even be assigned to an entity that cannot be held accountable for its actions? If an AI causes damage, how can it be punished? These issues are raised by many law scholars when dealing with the liability assignment of an act with legal consequences by an AI. At present, liability usually goes towards the manufacturer or user of an AI, so the AI system itself cannot be seen as an agent.

But Should AI Be Considered an Agent?

As I have argued above, depending on how you define agency, the idea of AI being an agent can be seen as either reasonable or completely absurd. Given that it is a possibility, should we consider AI as an agent? Even though we often derive whether an entity should receive any consideration from its ontology and capabilities, should we apply the same reasoning when dealing with AI? Would that be beneficial to our society, our legal systems, or even humanity as a whole? Should we even ask that question?

With the fast development of AI, we keep dwelling on what each system can and cannot do; we thereby neglect the question of whether this consideration is the right one to focus on. What if, instead of focusing on what an AI can do, we center discussion on whether these entities can be seen as agents no matter how complex, intelligent, or autonomous they might be? Although the abilities and inabilities of current AI systems are important to the discussion of the position of AI in society, this might better be left as a follow-up question to the most immediate inquiry: given the lack of agreement on the definition of agency and regardless of the abilities of these newly developed entities, is it socially beneficial or possible to consider AIs as

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

agents?

6. How does AI affect human Autonomy?

Carina Prunkl
Senior Research Scholar, Future of Humanity Institute, University of Oxford

Autonomy (*autos* = self; *nomos* = law) in the context of human beings refers to the capacity of self-governance or self-determination. This also implies that an individual's actions are neither the product of external manipulation, nor imposition of external forces. Autonomy in this sense plays an important role in Western culture and is often considered desirable for the individual. When we speak about 'autonomous systems' in the context of artificial intelligence, we similarly refer to some sort of 'self-governance', but in contrast to human case, this 'autonomy' has little to do with acting true to one's own beliefs, desires or motivations. Instead, it refers to the capacity of the system to learn and perform certain tasks without human guidance or supervision. A well-known example of such 'autonomous systems' are self-driving cars that navigate themselves through traffic to bring their passengers from A to B. But of course this type of 'autonomy' is not limited to the mechanical realm and we may easily conceive of virtual 'autonomous systems', such as virtual assistants that organize our lives by making appointments, doing (online) grocery shopping, taking notes, etc. By outsourcing seemingly trivial tasks such as driving and grocery shopping – not to mention some highly non-trivial tasks, such as those now performed by soldiers but that might at some point become automated – we are handing over more and more responsibilities to 'autonomous systems.' How will such a development affect our own autonomy? It is difficult to imagine that at least those of us who are somewhat indifferent to the joy of driving, will feel or be

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

less autonomous by having a car that takes us to where we want to go faster and safer. This is at least in part because it is we, after all, who decide where to go and when. But what about when such 'autonomous systems' not only navigate us through traffic, but also through life? When they learn from our behavioral patterns, our preferences, our relationships, to make predictions about, say, what groceries we would like to eat next week? Here the situation is much less clear. Do we gain autonomy by not having to be bothered with boring grocery planning and shopping, and instead having time for the things we would really like to do? Or do we instead forfeit autonomy by not being the ones who make the choices about our nutrition, returning almost to the childlike state of not having to take responsibility for certain aspects of our lives? These are questions we urgently need to ask ourselves.

7. The Myth of Agency

Sarah Newman
Senior Researcher, Principal at metaLAB at Harvard
Fellow, Berkman Klein Center for Internet & Society, Harvard University

"Ultimately, nothing or almost nothing about what a person does seems to be under his control."
– Thomas Nagel, Moral Luck

We look, critically, at how our technologies work, and yet we make assumptions about how we work. What motivates our choices? Are we in control of our actions – and if so, all of them, or only some of them? As our interactions with and dependence on new technologies, including AI, become both increasingly common and invisible, what, if any, agency are we giving up? If we better understand our agency, how does this connect to our responsibility for the technological world we

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

are creating, and the natural world we are destroying? What responsibilities should we have for our own behaviors, and where does accountability reside in automated systems?

We use the term "agency" to refer to humans, to current and future AI systems, as part of a framework for responsibility and accountability. But what do we mean by agency? Agency is defined differently across disciplines–from computer science to philosophy to sociology to law. Recent developments in neuroscience and AI both call into question the accepted notion of volitional agency as the willed proximate cause of a thought or an action. How might exploring frameworks of agency affect our approaches to ethical standards in the development of AI? A potential blind spot in our analysis of the development of AI lies in the assumptions we make about our own agency, freedom of will, and moral capabilities.

Are we are actually more accurate when describing the behavior of machines–mechanistic, physical, governed by the laws of nature and programming–than we are when we describe ourselves? Things get fuzzy as the mysteries of consciousness and subjectivity arise. What is true – and what, if not true, is useful to believe?

We believe that we, as humans, have at least some agency. We acknowledge that our degrees of agency differ across individuals and circumstances, increasing or decreasing based on certain constraints, and governed by physical laws–at least those outside of our brains. Most people don't believe that they could defy physical laws: the laws of gravity, survival without food, etc. We accept these physical constraints, those that appear to affect all beings and appear to be external to us, or at least external to our physical bodies. Yes, this agency is highly variable: a healthy adult has more agency, people tend to agree, than a baby, or someone who is very

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

old or unwell.

We tend to agree that we do not have the agency to fly, or to travel in time, or countless other fantastical things (barring of course certain mental illnesses, or other illnesses which impinge on mental capacities, which have their own unique relationship to agency and thus also to responsibility). And yet most people now, and throughout history–across cultures, ages, and every other demographic factor–have a distinct sense of *being in control* of (at least some of) our behaviors and actions. Even though it is difficult to explain, there is a distinct and overwhelming sense that I am *choosing* to write these words, that I will *choose* what to have for dinner, that I could choose to clap my hands, or nod my head, or close my eyes. This sense, as inexplicable, biologically and physically, as it may be (as a being comprised of physical matter that came into existence in a way that I certainly did not *will*), from where did it arise? *Is the sense of agency I possess merely a myth?* Perhaps a useful, or even inescapable myth? If so, is considering such questions useful or productive?

For me, reflecting on such questions is enriching: it enriches my daily life and my experiences. Paying attention to this deep and abiding mystery, somewhat ironically, feels empowering – as if I am curiously contemplating whether the backdrop is a facade, whether this sense of agency is indeed an illusion. I acknowledge the possible privilege of this perspective. Perhaps, if I do indeed have some sort of inexplicable agency, contemplating it is enjoyable because I have (if I have it at all) a relatively high degree of it. But perhaps not.

Such topics have fascinated philosophers, theologists, and most humans for so long as we have records of such contemplation. Debates on free will or the existence of agency–nevertheless have barely made their way into discussions of the new sophisticated technologies we are creating–particularly AI, in terms of how it already

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

is acting in the world, as well as how it could impact the future. We talk about autonomy and responsibility, but can we use this moment to also reflect back on our assumptions about ourselves?

1. A potentially defensible extension might argue that moral responsibility can only be rooted in agents that have some ability to modify their target goals.
2. The Book of Job is a clear example of this dynamic. The work establishes repeatedly that Job was blameless. Yet, three of Job's four friends insist that Job's misfortune is *just* punishment for his sins, as God is necessarily just.
3. Max Weber (1919) "Politics as Vocation"
4. There is an implicit assumption here: that all reality/any observable phenomenon is rational and discoverable.
5. Birhane, A. (2017). Descartes Was Wrong: 'A Person Is a Person through Other Persons'. *Aeon*.
6. Weiser, M. (1995, June). The computer for the 21st century. In Human-computer interaction (pp. 933-940). Morgan Kaufmann Publishers Inc.
7. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* Profile Books.
8. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.
9. Bowles, N (2018) https://www.nytimes.com/2018/10/26/style/digital-divide-screens-schools.html
10. Bowles, N (2018) https://www.nytimes.com/2018/10/26/style/silicon-valley-nannies.html
11. Margaret S. Archer, "Morphogenesis versus Structuration: On Combining Structure and Action," *The British Journal of Sociology* 61, no. s1 (2010): 225–52, https://doi.org/10.1111/j.1468-4446.2009.01245.x.

12. Douglas V. Porpora, *Reconstructing Sociology: The Critical Realist Approach* (Cambridge: Cambridge University Press, 2016).
13. Bruno Latour, "On Actor-Network Theory: A Few Clarifications," *Soziale Welt* 47, no. 4 (January 1, 1996): 369–81, https://doi.org/10.2307/40878163.
14. Edwin Sayes, "Actor–Network Theory and Methodology: Just What Does It Mean to Say That Nonhumans Have Agency?," *Social Studies of Science* 44, no. 1 (2014): 134–49.
15. Bruno Latour, "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts," in *Shaping Technology/Building Society: Studies in Sociotechnical Change*, ed. Wiebe E. Bijker and John Law (Cambridge, MA: MIT Press, 1992), 225–58.
16. Michel Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay," *The Sociological Review* 32 (1984): 196–233, https://doi.org/10.1111/j.1467-954X.1984.tb00113.x.
17. Catriona Mackenzie and Natalie Stoljar, eds., *Relational Autonomy: Feminist Perspectives on Automony, Agency, and the Social Self* (Oxford: Oxford University Press, 2000), http://ebookcentral.proquest.com/lib/ubc/detail.action?docID=430598.
18. Jason Edward Lewis et al., "Making Kin with the Machines," *Journal of Design and Science*, July 16, 2018, https://doi.org/10.21428/bfafd97b.
19. Jerry Lee Rosiek, Jimmy Snyder, and Scott L. Pratt, "The New Materialisms and Indigenous Theories of Non-Human Agency: Making the Case for Respectful Anti-Colonial Engagement," *Qualitative Inquiry*, February 27, 2019, 1077800419830135, https://doi.org/10.1177/1077800419830135.
20. In Barad's "agential realism", drawing on Latour and quantum theory, "agency is not an attribute but the ongoing reconfigurings of the world. The universe is agential intra-activity in its becoming". Karen Barad, *Meeting the Universe Halfway: Quantum*

by: Sarah Newman, Abeba Birhane, Mike Zajko, Osonde A. Osoba, Carina Prunkl, Gabriel Lima, Jon Bowen, Rich Sutton and Cathy Adams

Physics and the Entanglement of Matter and Meaning (Durham: Duke University Press, 2007), 141