Agents without Agency: Anthropological and Sociological Lessons for Contemporary AI Research and Policy

Greta Timaite, Michael Castelle

Centre for Interdisciplinary Methodologies
University of Warwick
G.Timaite@warwick.ac.uk, M.Castelle.1@warwick.ac.uk

Abstract

Recent hype about artificial neural network-based agents, fomented by tech corporations and AI startups and increasingly parroted by public sector actors and even critics, is premised on the idea that large deep learning models can act more and more autonomously — with minimal or no human supervision — and/or with "increasing agency". In this paper, we contest the assumption — embedded in decades of research on the topic of agency in various AI subfields — that agency should be understood purely as a property of an individual. To challenge these claims, we will combine a historical analysis of past AI research with evidence from social-scientific scholarship, specifically from empirically-grounded sociological theory and linguistic anthropology, to make a case for a conception of agency that is ontologically relational and inextricable from social accountability. "Agency" is thus revealed as not a locatable and/or quantifiable property of an individual subject, but as a determination emerging from social interactions, whose status can dynamically vary with the observer and/or the accompanying sociotechnical (and/or legal) context. This perspective can a) illustrate the potential incentive to convince the public of the "agentic" status of AI models, which can defer accountability and blame away from corporations and/or developers and onto instances of opaque neural network architectures; and b) provide empirical grounding for the regulatory "social licensing" of purported agentic models.

1 Introduction

In recent years, a particular technological discourse has gained traction: the idea that artificial intelligence (AI) *agents* are the next frontier of AI (Dhamodharan 2025). For example, in 2024, Google introduced Gemini 2.0 as a model for the "new agentic era"; and Sam Altman, the CEO of OpenAI, recently claimed that AI agents joining the workforce as virtual employees would lead to "great, broadly-distributed outcomes" (2025). The hype about the potential of AI agents also comes from international organizations such as the World Economic Forum (Larsen and Li 2024) and university-based researchers (Park et al. 2023; Xi et al. 2023; Durante et al. 2024). We will argue here, using evidence from sociology and anthropology, that a) the common view that agency is an individual *property* of an AI model is incoherent, because agency is (across a wide variety of human cultures) a fundamentally *social* (or *relational*) concept; and that b) the claim that such (fictitious and/or ill-defined) property-like agency is "increasing" is equally nonsensical. We will also argue c) that such rhetoric is conveniently advantageous for actors who may in the near future seek to avoid blame for the actions of so-called "agents" that they produce and distribute.

For example, Alan Chan and colleagues (2023) have described the ethical concerns and potential long- and short-term harms that could be brought about by AI agents. However, they avoid defining agency directly and instead refer to a concept of "increasing agency" of an algorithmic system, characterized by a mix of observable and unobservable properties; this work has been subsequently cited and mirrored by other commentaries on the ethics and potential harms of agents from industry figures at DeepMind and OpenAI (Manzini et al. 2024; Shavit et al. 2023; Gabriel et al. 2024). The concept of agency used by these scholars and companies, as well as across much of the history of mainstream AI in both the late 20th century and in its current connectionist revival, is profoundly *individualist*: agency is unproblematically envisioned as something that a subject possessively "has" (to greater or lesser extents), independent of their social (or sociotechnical) contexts. Such a perspective befits a discipline like computer science with strong historical connections to cognitive science, but in the social sciences debates about action, agency, and autonomy have been ongoing for some time, and the views which are most empirically aligned with observational/interactional studies are ones in which "agency" is not a property or set of properties but an (overt or covert) *attribution* of one subject by another. For example, we can contrast these recent claims about the increasing agency of AI systems to a comment made by a linguistic anthropologist Laura Ahearn over two decades ago that "[i]t is *not useful... to talk of having "more," "less," or even "no" agency... agency is not a quantity that can be measured.* Rather, researchers should focus on... different ways in which agency is socioculturally mediated in particular times and places" (2001, p. 122, emphasis added). And in recent work synthesizing and improving upon historical debates about agency in the social sciences, the sociological theorist Fabian Anicker and colleagues write: "[i]nstead of asking what agency "really" is, we are led to ask: How do people distinguish between agents and non-agents, and

under which conditions do people start treating AI systems as agents?" (Anicker, Flaßhoff, and Marcinkowski 2024, p. 312, emphasis added). This latter question — of trying to understand not how agency emerges in the individual, but under what conditions humans *attribute* agency to other (human or non-human) subjects — indeed recalls an important intervention at the very origins of the field of AI, namely Alan Turing's 'Test' for intelligence (Turing 1950).

We can see (whether via Turing or Anicker's work) that the phenomena of concern is not the presence or absence of some (greater or lesser) property of agency in an individual entity, but instead the *ascription* of agency to such entities by others — and this is indeed a phenomena that can be seen to be increasing in the wake of the widespread deployment of generative large language models. For example, in a recent study, LLM chatbots for some human interlocutors take on an agentic quality (as expressed in a post-experiment questionnaire); but for others, they do not (Anicker and Flaßhoff 2024). Just as with Turing's "intelligence", humans can differ (for various reasons) on whether a subject should be considered agentic, even among those with a shared language and culture. Ahearn, by contrast, is inspired by anthropological studies of certain non-Western cultures and languages in which a speaker's assessment of a subject's agency can (or must) be overtly marked in the grammar of their speech.[1] Such obligatory and empirically observable markers provide anthropologists with accessible data on who and what is considered "agent-like" by who and in what types of social situations. Her quote above is critiquing what might be called the *commensuration* of agency — the transformation of "qualities into quantities, difference into magnitude" (Espeland and Stevens 1998, p. 316). Such transformations are perhaps natural to those embedded in the epistemic community of AI in which "intelligent" models are, of course, trained entirely with reference to a unidimensonal loss function; but we hope to clarify here that the perception of "increasingly agentic" contemporary AI models is ultimately not simply a function of their weight parameters.

This creates an occasion to ask: why do AI researchers (and some social scientists) find it useful to talk about degrees of agency in individual AI models/systems? What kind of assumptions underpin such claims? And how would these claims change under a more socially coherent conception of agency? To answer this, we examine the historical and contemporary concept of agency in AI research. Specifically, we ask how AI researchers *talk* about AI agency in published research documents (see section 4). We will see that the various discursive claims about non-human agency in both 20th and 21st-century AI research can reveal something about the self-perceived agency — and, inexorably, accountability — of the researchers, institutions, and corporations which have tasked themselves with bringing such "agents" into existence.

[1]For linguists, speakers of languages with so-called *ergative case marking* explicitly mark the active participant of transitive verbs differently (such as "the boy" in "the boy threw the rock") from the (sole) participant of intransitive verbs (such as "the boy" in "the boy fell") (Dixon 1979); e.g., the former might be encoded with some ergative suffix/particle: "boy-ERG threw rock-ABS", while the latter would use a differing absolutive suffix/particle: "boy-ABS fell". Such languages can sometimes — as in the case of Samoan as studied by Duranti (1994) — provide overt grammatical clues to the perception of a subject's agency, as opposed to languages like English which can easily downplay or obscure agency with e.g. passive-voice constructions.

2 Related Work

2.1 Social Embeddedness of AI

Social researchers have long argued that technology intervenes, shapes, and represents social life (Bijker and Law 1994) — and contemporary AI technologies are no exception (Roberge and Castelle 2021). For example, Langdon Winner's classic text "Do Artifacts Have Politics?" is often cited to illustrate how technology can establish and/or maintain forms of social order as a consequence of it being embedded in social life (Winner 1980). The significance of this text for the AI ethics/society community is reflected in a work by Scheuerman and colleagues (2021), "Do Datasets Have Politics?", in which datasets in computer vision research are approached as value-laden artifacts. They demonstrate that these artifacts, contrary to popular rhetoric, possess a fundamental "interpretative flexibility" (Pinch and Bijker 1984) — meaning that such vision datasets can be designed/interpreted in more than one way because they are social and cultural products.

In the case of contemporary AI, the question of representation of various (often minority) social groups has received much attention. Researchers argue that categories of identity, such as gender or race, are not inherent and fixed characteristics of individual subjects, but rather social constructs that vary across time and space (Gebru 2020; Benthall and Haynes 2019; Lingel and Crawford 2020). Recognition that AI has the potential to cause harm through its growing presence in social life (Shelby et al. 2023) has led researchers to suggest various strategies to push against categorical simplification in AI research (Buolamwini and Gebru 2018; Benthall and Haynes 2019; Devinney, Björklund, and Björklund 2022; Gadiraju et al. 2023; Qadri et al. 2023).

There is also another strand of research calling for a critical reflection on AI research concepts. Recently, the concept of "representation" in AI research was studied using the case of evaluation datasets (Bergman et al. 2023), suggesting that in order to account for the historical situatedness of datasets, it is important to ask not only who or what is represented, but also *when*. Other researchers argue that it remains important to attend to the social construction of "intelligence" and to examine the relation between ethical and techno-scientific dimensions of machine learning (ML) benchmarks (Blili-Hamelin and Hancox-Li 2023).[2] The Turing Test, which is most often taken as *the* test to determine AI's "intelligence", is frequently invoked to make claims about the progress of AI research even though it lacks a standard description (Mitchell 2024). Despite this — or perhaps

[2]The questions about intelligence, knowledge, and representation in AI research are not new as these topics have been explored in the context of symbolic AI, a dominant AI research paradigm in the late 20th century (Adam 1998; Dreyfus 1992; Hunter 1999).

because of this — there have been recent proclamations that the Turing Test has been "won" because, statistically, chatbots exhibit behavior similar to humans (Mei et al. 2024; Scott 2024); yet sociologists have long argued that claims about AI's "intelligence" can only be evaluated with respect to the social context in which it is embedded because (as we noted in the introduction) "intelligence" is a category which depends on ascription by an observer (Collins 1990).

The overarching aim of these critiques of AI is to demonstrate and situate AI as a product of designers' social and cultural milieu. In other words, AI is not neutral: values, ideas, and ideologies are represented in technologies that, when technology is embedded in social life, might lead to biased outcomes for some social groups. We will argue, however, that it is time to extend critical engagement with AI beyond critiques of bias and "intelligence": we need to move towards better understanding of how AI, through its increasingly ubiquitous participation in everyday life, problematizes the very concepts used in researching and understanding social life.

2.2 AGI and Autonomy

Blili-Hamelin et al. (2025) has recently cautioned against framing "artificial general intelligence" (AGI) as the primary goal of AI research. Not only, they argue, is it not clear what counts as AGI, but its two primary conceptual building blocks — *intelligence* and *generality* — are also contested. Gebru and Torres (2024), too, argue that AI researchers should move away from the goal to build "all-knowing" systems that cannot be tested for safety. They suggest, moreover, that AGI discourse — intellectually rooted in early-20th century utilitarianism and eugenics — is camouflaged by the rhetoric of AI safety: companies can simultaneously make claims about importance of AI safety while developing unsafe products and evading accountability. These critiques of AGI have rightly focused on the notion of "intelligence" but have yet to pay sufficient attention to another (and increasingly significant) concept in contemporary AGI discourse, namely that of *agency* and/or *agents*.

Recently, there have been various attempts in the AI research community to operationalise an A(G)I agent. To do so, the notion of *autonomy* is frequently used as a substitute for "agency" in characterising an agent. For example, Morris et al. (2024) suggest autonomy to be one of the quantifiable attributes, alongside generality and performance, in operationalizing AGI; Mitchell et al. (2025), in their recent position paper on "fully autonomous" AI agents, split agency and autonomy to circumvent philosophical debates about *intentionality* of actions (largely derived from the analytic tradition, which we will discuss in the next subsection). Arguably, both works characterise autonomy of an AI agent relationally in so far as "autonomy levels" (Morris et al. 2024) or "agentic levels" (Mitchell et al. 2025) are related to how much control a human actor exercises over an "agentic" system; however, the authors succumb to the individualistic assumption that "full autonomy" is an attainable property. Although this "turn to autonomy" is nothing new in AI research (as we will see in section 4), both STS scholars and linguistic anthropologists have found the concept of "autonomy" problematic and ultimately inextricable from its origins in the Enlightenment traditions of individual liberty and moral judgment (Suchman and Weber 2016; Duranti 2015).

2.3 Technology, Agency, and Society

In this subsection, we will engage with two distinct late-20th century intellectual approaches to agency outside of AI and computer science research that have been used to examine and/or study technological artifacts as agents — one deriving from analytical philosophy and the other deriving from the subfield of actor-network theory (ANT) within science and technology studies (STS). By contrast, we will argue that neither is sufficient to help us understand contemporary claims about "agentic" AI, as they either operate with a strongly individualist notion of agency or they fail to address questions of power and accountability. In light of these limitations, we will briefly discuss another approach which we think is more useful for developing a better understanding of claims about AI agency — one that frames agency as a matter of situational (and ontologically *social*) attribution in which the question is fundamentally an empirical one: *how does an entity become an agent?*

"Agency" as Intentionality In analytical philosophy, the notion of *intentionality* and various adjacent expressions, such as "intent" or "intentional act", are invoked in discussions of agency (Davidson 1971). John Searle's so-called "Chinese Room" thought experiment (Searle 1980) is a relevant case, in which an English-speaking human agent produces correct sentences in the Chinese language with the help of instructions written in English that explain how to do so. However, Searle argues that just because a human agent can provide correct outputs in Chinese, it does not mean that he understands Chinese and, therefore, rule-driven symbol manipulation should not be equated to understanding. The only machine that could think, according to Searle (1980, p. 423), is that which has "the same causal powers as brains"; and these causal powers require what is called "intentionality" (Searle 1979) — a concept introduced by Brentano (1995 [1874]) — defined as the individual's capacity to direct mental states to objects in the world. Intentionality, for Searle (2016), has a biological origin in which the meaning of linguistic expressions depend on underlying propositionally-structured mental states.[3]

In general, there are at least two problems with a conceptualization of agency that draws from this analytic-philosophical tradition. First, the category of "human" as an analytical category is often approached unproblematically. This is one source of contention for critics like Symons and Abumusab (2024), who note that the views of agency informed by "human adult-level cognition" risk excluding some social groups, e.g. children. Instead, they advocate for a minimal account of agency that attends to the context and does not presume that an agent has "its own intentions, mental representations, etc." (2024, p. 16). Second, it is often

[3]More recently, a similar line of argument has been advanced by linguists Bender and Koller (2020) in response to claims about the ability of LLMs to understand. The analytical problems discussed in this section also apply to their work.

assumed that agency (dependent as it is on some individual "communicative intent") is a property of an individual. While Symons and Abumusab attempt to conceptualize "artificial agency" as multidimensional and context-sensitive, it is still something property-like that an individual entity "has", and therefore their work implicitly shares assumptions made by the the analytic philosophers.

"Agency" as an Effect The most (in)famous argument for including non-humans in social research was advanced by Bruno Latour and other proponents of so-called Actor-Network Theory (ANT). Latour noted that one of the reasons nonhumans have been excluded from social research is due to a limited notion of agency. He wrote: "If action is limited a priori to what 'intentional', 'meaningful' humans do, it is hard to see how a hammer, a basket, a door closer, a cat, a rug, a mug, a list, or a tag could act" (Latour 2005, p. 71). For Latour, and ANT-inspired scholars, an agent (or in the ANT nomenclature an *actor*) is that which makes a difference, thus turning the question of agency into an empirical problem (de Laet and Mol 2000; Sayes 2014). By adapting a minimalist notion of agency, ANT moved away from the "undecidable" metaphysical debates in traditional sociology between the primacy of "structure" vs. "agency" (Emirbayer and Mische 1998) and moved towards an understanding of agency as an *effect* (Callon and Law 1997).[4] Approaching agency as an effect decouples it from the assumption that agency is localised in an individual subject, an assumption that, as STS scholar Andrew Pickering (2024) reminds us, is contrary to much philosophical thought that attempts to ascribe "agentic" attributes to individuals.

An Alternative: "Agency" as Social License While a minimalist notion of agency makes it possible for ANT scholars to study agency of non-humans in social research, ANT — akin to analytic philosophy — largely leaves the categories of "human" and "nonhuman" unexamined. As feminist sociologist Monica Casper (1994) argued, the question of social *attribution* of agency is not only a significant empirical question for understanding the construction of humanity but also a political one: for example, the attribution of agency to non-humans (as is common in ANT scholarship) might serve to deflect human accountability and power in structuring social relations.

Echoing Casper's argument in the context of autonomous weapons systems, Suchman and Weber (2016) argued for the need to articulate human-machine differences while also advocating for the reconceptualization of agency and autonomy from an individual understanding to a sociotechnical and relational one. They propose a "shift in conceptions of agency and autonomy, from attributes inherent in entities to effects of discourses and material practices that either conjoin humans and machines or delineate differences between them" (Suchman and Weber 2016, p. 76), and reject "the premise that autonomy can be adequately understood as being an intrinsic capacity of an entity" (Suchman and Weber 2016, p. 78).

More recently, responding to the theoretical challenges posed by AI to sociological research, Anicker and colleagues (2024) proposed a theory of AI agency that takes agency as an empirical question: i.e., we should ask and study how humans come to treat AI as "an agent" who can be held accountable for its actions. In other words, in line with Casper and Suchman/Weber, they are interested in the attribution of agency, or to use their vocabulary, the *social licensing* of agency.[5] It is this sociological approach to agency, we will argue, that can be usefully put in conversation with theories and methods developed in linguistic anthropology to help us make sense of both historical and contemporary AI agent discourse.

3 AI Document Collection and Analysis

Drawing on interpretative approaches, we take individual AI research publications as our unit of analysis to explore what AI researchers from the mid-1990s to the present say about agency in their work. Analysis of scientific documents is a well-established method in social research about AI. For example, quantitative methods have been used in AI research evaluation (Klinger, Mateos-Garcia, and Stathoulopoulos 2022) and to examine the role and impact of Big Tech in shaping AI research (Ahmed and Wahed 2020). Others have combined quantitative scientometric methods with interpretative approaches to study the discursive role of algorithms in AI research (Munk et al. 2024). Qualitative methods, such as close reading of a small number of selected texts, are also a common empirical strategy in interpretative sociological and anthropological research to study knowledge construction (Latour and Woolgar 1986), discursive practices (Amoore et al. 2023), and agency (Neff and Nagy 2016; Ahearn 2010). Relevantly for our work, Ahearn writes that "[one] way of analyzing agency in [anthropological studies of] language is to look for how people talk about agency — how they talk about their own actions and others' actions, how they attribute responsibility for events, [and] how they describe their own and others' decision-making processes" (Ahearn 2010, p. 41).

In the context of AI research, different identification strategies have been developed to select relevant publications. For example, some rely on field categorizations provided by academic databases (Frank et al. 2019) or AI-centered vocabularies to select relevant keywords (Gargiulo et al. 2023), while others consult AI experts to select relevant keywords for database querying (Ahmed and Wahed 2020). We used an iterative approach to collect publications that combines database querying with a snowballing technique. A similar approach to data collection has been used to examine values (Birhane et al. 2022), assumptions about

[4]The conceptualization of agency as an effect is not most commonly associated with, but not limited, to ANT literature. Relational sociologists, for example, also discuss agency as effect (Burkitt 2016).

[5]The vocabulary of "social licensing" of agency here is derived independently from the "principal-agent" perspective of agency in law and economics (Watts and Reynolds 2021). While both approaches imply relationality in so far that agency emerges between at least two actors, the former is more general while the latter specifically presumes a relationship of *delegation* (Shapiro 2005).

gender (Devinney, Björklund, and Björklund 2022) and expertise (Diaz and Smith 2024) in AI research. We started by querying Google Scholar with a list of keywords, such as "AI (+) agent" and "language agent", to select and carefully examine the most well-cited works to find out how an agent and/or agency is defined in a publication. The so-called snowballing approach (Wohlin 2014) was used to expand our corpus of publications: for each candidate publication we examined, if applicable, the publications cited to define agency (backward snowballing) and, in addition, we also reviewed and read various publications citing the publication in question to understand that definition's influence in the AI community (forward sampling).

To limit the scope of analysis, we selected a small number of definitions of agency in AI research that could be considered exemplary or "ideal-type" (Weber 2012 [1904]). Some of these selected definitions are also present in other recent works examining agency in AI research (Kasirzadeh and Gabriel 2025; Mitchell et al. 2025), but in our work we also paid attention to definitions of agency coming from some AI-adjacent fields, such as robotics, due to the multidisciplinary roots of AI agent research (see section 5).

3.1 Evaluative Grid

With inspiration from empirical social research studying AI controversies, we adapted an *evaluative grid* (Marres et al. 2024) as an analytical strategy to help articulate implicit and explicit assumptions. Moreover, we use this evaluative grid, represented in Table 1, as a visual aid to a) demonstrate the overall conceptual incoherence of agency in AI research, and to b) contrast it against more ontologically social conceptions of agency. It should be noted that our aim is not to construct a complete taxonomy of AI agents past and present, but to evaluate and critically comment on agent/agency definitions. In this respect, our work differs from Kasirzadeh and Gabriel (2025) who rely uncritically on definitions coming from AI research to develop a more holistic framework that still retains those definitions' assumptions.

To attempt to resolve these conflicting and often *ad hoc* definitions of AI agency, we adopt a much broader and more culturally universal conception of agency from linguistic anthropology, specifically the work of Enfield and Kockelman on what they call *distributed agency*, which considers agency as a multi-dimensional and social phenomenon which can primarily be described along (at least) two dimensions: *flexibility* and *accountability* (Enfield and Kockelman 2017, pp. 4-7). Enfield (2013) defines *flexibility* in terms of a threefold capacity (whether individual or sociotechnically "distributed"):

- *control*, or ability to determine the occurrence of perceptible behavior;
- *composition*, or capacity to design one's behavior to accomplish something;
- *anticipation/subprehension*, or capacity to foresee the reaction of others to one's action.

In other words, flexibility is about the capacity to determine behavior and outcomes. *Accountability*, on the other hand, refers to how others interpret and act on the above observed behavior, and can then be further specified into three categories:

- *evaluation*, such as having one's behavior praised or blamed, and to give reasons for said behavior on demand;
- *entitlement*, or a right, invoked by an agent or others, to carry out certain behavior and to give reasons for it;
- *obligation*, or a duty, invoked by an agent or others, to carry out certain behavior and to give reasons for it.

We will return to Enfield and Kockelman's (2017) definition of accountability later in the paper but, generally, we find it helpful to think about accountability as answerability (Nissenbaum 1996; Kroll 2020). This is because accountability is a social/relational phenomenon, which cannot be fully located within any isolated "agentic" individual.[6] For example, an individual's action can be evaluated as violating social norms and, therefore, they might be *blamed* and expected to answer for inappropriate behaviour (and such blame might or might not be legally sanctioned). Given this discussion, we define the three main sections of our evaluative grid:

- *Unit of analysis* is a parameter that refers to the assumption of where agency is located. Agency can be understood as either:
  - *Individual*, meaning that it is studied as an individual property;
  - *Distributed* among multiple human and/or nonhuman agents.[7]
- *Flexibility* is a set of parameters that refer to the assumption about an agent's capacity for action and anticipation of its outcomes. Our aim is not to detail an agent's flexibility but to ask whether an element of flexibility is present in a given definition of agency. Thus, each aforementioned element of Enfield and Kockelman's conception of flexibility – *control*, *composition*, and *anticipation* – in Table 1 is marked as either present (✓) or not (✗).
- *Accountability* is a set of parameters that indicate which, if any, elements of social accountability or answerability, namely *evaluation*, *entitlement*, and *obligation*, is assumed to be a dimension of agency (✓) or not (✗).

[6] Our conception of agency, while depending on the intrinsically relational concept of *accountability*, does not depend on the concept of *responsibility*, a term which is commonly conflated with accountability but which is often, but not always, imbricated in an Anglophone liberal tradition with notions of individual or "personal" morality (Nissenbaum 1996; Bovens 2007; Coeckelbergh 2023; Lechterman 2024; Hill and Irvine 1993). However, there have been attempts to bring such individualist moral philosophy together with relational approaches towards accountability in the context of AI (Cooper et al. 2022).

[7] To understand the idea of "distributed" agency, consider 1) a human walking, 2) a human driving a car and 3) a human behind the wheel of a "self-driving" car, designed and manufactured by an external corporation; in both cases the aspects of flexibility and accountability vary, unsurprisingly leading to various debates about who is answerable in the case of accidents involving self-driving cars (Stilgoe 2018; Marres et al. 2025).

4 Agency in Early AI Agent Research

In this section, we revisit earlier agent research, drawing on the aforementioned evaluative grid, to a) establish that certain concepts of agency are largely devoid of reference to social and/or relational accountability and b) to inform and contextualize our later discussion about contemporary claims about neural network-based AI agents.

4.1 The Rise of AI Agents: Technological and Critical Views

Oliver G. Selfridge's "Pandemonium" (1958) is often claimed to be the publication in which the concept of an agent was "coined" or "introduced" to AI research (Kay 1984; Lieberman and Selker 2003). Nonetheless, the agent concept was rarely an explicit focus of interest within the AI community until the 1990s, when it became taken seriously in academic and industry research (Nwana and Ndumu 1999; Wooldridge and Jennings 1995). Why was this the case? We will offer several insights into this question by examining internal accounts of the history of AI research.

Until the 1970s, according to AI/robotics pioneer Nils Nilsson (1995), the AI community directed its attention to problem-solving tasks, often involving explicitly symbolic rule-based systems, exemplified by projects like Allen Newell's Soar architecture (Laird, Newell, and Rosenbloom 1987). Newell's own intellectual history of AI suggests that this focus was largely a consequence of an intellectual divide that occurred in the late 1950s between the two approaches to building "intelligent" machines: proponents of symbolic systems focused on problem-solving tasks, such as playing chess, while the supporters of "continuous systems" (including early artificial neural networks such as Rosenblatt (1958)'s Perceptron) concentrated on pattern recognition as a central problem to solve (Newell 1982).[8] It is through the field of robotics, Newell (1982) argued, that the problem of recognition re-entered AI research, such as in Nilsson's work on mobile robot Shakey (Nilsson 1984), but it remained largely secondary to symbolic problem-solving and planning.

In the late 1970s and 1980s, AI research, motivated by external pressures to demonstrate its practical usefulness, moved away from toy problems in pursuit of generality to domain-specific (but still symbolic-centric) applications, notably expert systems (Nilsson 1995; Newell 1982).[9] In the 1980s, moreover, there were two intellectual developments that were significant in the rise and development of agent research. First, a subfield of "Distributed AI" (DAI) emerged that approached intelligence as a distributed accomplishment rather than a property of an individual (Davis 1980; Gasser 1991) and laid the foundation, in particular with its branch on *multi-agent* research, for research in the 1990s on "software agents" (Nwana 1996). Second, a major development in the late 1980s / early 1990s occurred in the field of robotics: there was a shift from a "top-down" (symbolic AI) to a "bottom-up" (behavioral and often embodied) view of intelligence (Maes 1993). Rodney Brooks (1991b; 1991a) explained that the behavioral approach grew from dissatisfaction with the performance of robots in real-world situations as, at the time, they demanded a carefully engineered environment (as exemplified by Nilsson's Shakey robot). Instead, Brooks (1991b, p. 1227) proposed to build mobile robots by "considering the problems of building an autonomous agent" that could cope with dynamic changes in the environment without requiring recourse to symbolic reasoning. This was a "proof-of-concept" of the importance of situatedness and embodiment for genuine knowledge and/or intelligence, a position already taken by AI critics like Lucy Suchman (1987) and Hubert Dreyfus (1997). These two developments, which marked a change in how intelligent machines are to be built, can be considered a *critical* (or ideological) explanation for the subsequent rise of "agent"-centric AI research in the 1990s.

Finally, to fully appreciate the emergence of agents in AI research, we need to attend to the rise of certain communication technologies in the 1990s such as the World Wide Web which, from the outset, overwhelmed casual users with metaphoric "waves" of information. This practical problem led to suggestions from AI researchers that they should develop *agents* rather than *systems* to help handle information overload (Alonso 2002; Lieberman and Maulsby 1996; Nilsson 1995). These agents were no longer embodied and situated in a physical environment as Brooks' mobile robots but, rather, were virtual or software agents which could assist users in various computer-related tasks and, more generally, be treated as "personal assistants" (Maes 1994). Thus, the rise of digital communication technologies can be considered a *technological* explanation for the growing interest in agent-based view of AI; a shift which, Nils Nilsson (1995) argued, signaled a move — or rather a return — to AI's original goal of building general AI systems that might not have specialist knowledge programmed into them, but can instead use external tools to compensate for their limitations.

4.2 Individual and/or Asocial Autonomous Agents in the 1990s

Jeffrey Bradshaw, in a popular 1997 introduction to software agents, distinguished between *ascriptive* and *descriptive* understandings of agents (Bradshaw 1997). The "ascriptive" approach is grounded in the assumption that (as we have suggested above) agency cannot be reduced to a simple list of attributes, because the question of its agency is in the eye of its beholder — i.e., one person's intelligent agent can be someone else's dumb agent (or not an agent at all). The latter "descriptive" approach is characterised by the attribution of properties that make an individual entity agentic, such as the aforementioned "autonomy". It is this latter view which, according to Bradshaw, many agent researchers find more "acceptable" and, indeed, became prominent among researchers in AI and its adjacent fields.

[8] See Olazaran (1993) and Cardon, Cointet, and Mazières (2018) for historical analyses on the symbolic AI vs. connectionism controversy.

[9] There are some exceptions, such as Carl Hewitt's work on the so-called "Actor Model" of computation inspired by asynchronous communication processes (Hewitt 1977), but this work was more influential to distributed systems research than AI proper (Nwana 1996).

This preference for a descriptive approach to agency is illustrated by a well-cited article authored by multi-agent AI researchers Woodridge and Jennings as a response to the growing "noise" around agents in academic and industry research. They suggested two notions of agency: "weak" agency, characterised by autonomy, social ability, reactivity, and pro-activeness, and "strong" agency, underpinned by the aforementioned philosophical concept of intentionality (Wooldridge and Jennings 1995). Two years later, an article promisingly entitled "Is It an Agent, or Just a Program?" (Franklin and Graesser 1997) unfortunately also exemplifies a descriptive approach, assuming agency as a set of individual properties which defines an autonomous (software) agent using symbolic mathematical expressions. They claim that the agent's environment is a qualifying criterion: a program becomes an agent only "with respect to some environment" (p. 26). To understand the definitional importance of the notion of environment, we need to remember that, at the time, AI researchers focused on building expert systems and programs whose outputs did not explicitly depend on some external environment in which they were embedded.

To further evaluate and characterise the concept of agency in this era of AI research, we turn to the evaluative grid (Table 1). It is of little surprise that the majority of these agent definitions of 1990s AI research 1) imply a capacity for control and 2) assume agency to be an individual property. In the context of 1990s-era research, however, the idea of autonomy — which some declared to be the "missing ingredient" in AI research (Covrigaru and Lindsay 1991) — was more pertinent than that of "agency", and this was sometimes assumed to be self-explanatory (Franklin and Graesser 1997).[10]

For example, in their classic AI textbook, Russell and Norvig argued that autonomy is needed so that an ideal rational intelligent agent can operate in a dynamic environment (Russell and Norvig 1995). For them, "rational" behavior is not about the process of action selection but rather about a utilitarian notion of "doing the right thing" in order to accomplish a goal given the set of information that it has (Russell and Norvig 1995; Russell 1997). What, then, guides action selection to accomplish a goal? For Russell, reinforcement learning (RL) suggests an answer (and, as we will see in section 5, a societal challenge). Generally, through RL, an agent is trained to maximise a reward which can be either a) specified by a designer or b) obtained by observing others (as in the case of Russell (1998)'s "inverse RL"). For Russell and Norvig, an agent is "autonomous" if it has the ability to navigate an environment based on its previous experience (informed by a reward function) to accomplish a goal.

However, despite "autonomy" being a central notion in understanding agency, it is not conceptually coherent in AI research. Behavioral roboticist Tim Smithers (1997) remarked that such "undisciplined use" of the concept "robbed" the field by making it difficult to distinguish between autonomous and self-regulating agents. Smithers, who took inspiration from biologists such as Francisco Varela along with Heidegger's phenomenological work, defines autonomy as a self-steering behavior that originates from the self and cannot be predicted by an observer because it continually evolves (Smithers 1995, 1997). This definition contrasts with the more general one in Russell and Norvig's (1995) definition of autonomy: namely, a capacity to direct one's actions to achieve a goal. In this case, *who* defines *what* goal is not necessarily important in order for autonomy to be attributed to an agent. Therefore, despite conceptual inconsistency regarding the "origin" of goals guiding autonomous behaviour, both Smithers and Russell/Norvig assume autonomy to be an individual property. Multi-agent systems researcher Cristiano Castelfranchi (1994), on the other hand, offers a different and indeed "relational" understanding of autonomy, which for him should be understood not only by its relation to the environment (as in Russell/Norvig), but also in relation to its social context, or other agents. For Castelfranchi, the relational notion of autonomy involves not only anticipation of others' behavior (more explicitly discussed as "mind-reading" in Castelfranchi (1998)) but also accountability, such as an expectation to return a "favour" if an agent's actions "positively interfere" in another agent's world. Castelfranchi's conceptualisation of autonomous behaviour, which encompasses elements of flexibility and accountability, stands in contrast to other proposed definitions of AI agents, which effectively exclude accountability.

The evaluative task that we undertook in this section demonstrates that the notion of agency, as defined and debated in AI research in the 1990s, is semantically incoherent if we take relational accountability as fundamental to the concept of agency. Specifically, social relations are consistently bracketed out, leading to a conception of agency as a property (or properties) of an individual rather than, as we argue here, a situational determination that arises from social interactions and rests on (social and/or legal) accountability.[11] Significantly, this topic of accountability — about which AI researchers did not much concern themselves in the 1990s — has become a highly pertinent issue in contemporary discussions of the large language models that have, in recent years, begun to be referred to as "language agents".

5 Language Agents: Towards a Theory of Distributed Agency?

Contemporary claims about LLM-based AI agents can be interpreted along two intertwined discourses. First, there is what could be called a *pragmatic discourse* in which AI agents — LLMs equipped with function calls that can directly affect external reality — are framed as a potentially useful approach to solving technical challenges in, for instance, software development (Xia et al. 2024). At the time of writing, a vast amount of corporate labor (and concomitant hype) is devoted to implementing such language agents

[10]The definition of agency we use in this paper allows us to engage with what others mean by "autonomy" via the "flexibility" dimension (see section 4.1).

[11]This reductionist approach to the problem of accountability in agent research stands in a stark contrast to more socially oriented researchers who argued that accountability in "computerized societies" presents many challenges, notably the problem of "many hands" (Nissenbaum 1996).

	Source	Field	Agent	Autonomy	Unit	Flexibility			Accountability		
						con	*com*	*ant*	*evaluate*	*entitle*	*oblige*
1	Maes (1993)	software agents	"[a]n agent is a system that tries to fulfill a set of goals in a complex dynamic environment." (p. 2)	"[a]n agent is called autonomous if it operates completely autonomously, i.e. if it decides itself how to relate its sensor data to motor commands in such a way that its goals are attended to successfully." (pp. 2-3)	Individual	✓	✓	×	×	×	×
2	Castelfranchi (1994)	multi-agent systems	"a system whose behaviour is neither casual nor strictly causal, but teleonomic, "goal-oriented" toward a certain state of the world." (p. 57)	"[i]n a Multi-Agent world Autonomy is an intrinsically social notion" (p. 59)	Individual/ Distributed	✓	✓	✓	×	×	✓
3	Russell and Norvig (1995)	AI	"[a]n agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors." (p. 31)	"[a] truly autonomous intelligent agent should be able to operate successfully in a wide variety of environments, given sufficient time to adapt." (p. 35)	Individual	✓	✓	×	×	×	×
4	Smithers (1995)	robotics	"being an agent is a result of an ongoing interaction that has the right dynamics." (p. 152)	"[a]n agent is autonomous if it is able to cope with all the consequences of its actions to which it is subjected while remaining viable as a task-achieving agent in the world it operates in" (p. 123)	Individual	✓	✓	×	×	×	×
5	Wooldridge and Jennings (1995)	AI and multi-agent systems	an agent of "weak" agency has the properties of autonomy, social ability, reactivity, and pro-activeness whereas a "strong" notion of agency is applied to agents who exhibit human-like mental states (pp. 116-117)	"[autonomous] agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state" (p. 116)	Individual	✓	✓	×	×	×	×
6	Franklin and Graesser (1997)	AI and cognitive science	"[a]n autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to affect what it senses in the futures" (p. 25)	"[a]n agent that acts in pursuit of its own agenda is acting autonomously. It selects its own actions independently." (p. 26)	Individual	✓	✓	×	×	×	×

Table 1: Definitions of agency and autonomy in agent research in the mid-1990s.
(Under "Flexibility", "*con*" refers to *control*; "*com*" refers to *composition*; "*ant*" refers to *anticipation*.
Under "Accountability", *"evaluate"* refers to *evaluation*; *"entitle"* refers to *entitlement*; "*oblige*" refers to *"obligation"*.
See section 4.1 for definitions of the different elements of flexibility and accountability.)

characterized by their ability to directly and often asynchronously intervene in the world with the help of various function calls or "tools" (Wang et al. 2024). This pragmatic discourse about AI agents, while displaying a definitional legacy of the 1990s AI agent research, is not our focus in this section; but we will return to it in the conclusion. For us, a more urgent critical commentary is required to understand the *risk/safety discourse* around AI agents, often going under the rubric of "alignment", whose researchers constitute the so-called "AI safety epistemic community" (Ahmed et al. 2024) inside and outside of major AI tech companies.

One prominent early paper on the alignment of "Language Agents" published in 2021 (well before the introduction of ChatGPT) exemplifies the initial attempt to articulate the potential harms resulting from supposedly agentic LLMs (Kenton et al. 2021). Strongly influenced by discourse around "existential risk", they define language agents by contrasting them to speculative entities like "Oracle AI" (Armstrong, Sandberg, and Bostrom 2012): for them, language agents are not limited to question-answering capabil-

ities yet *are* restricted to text generation and cannot directly intervene in the world. Despite this, Kenton et al. (2021) argue, language agents can cause harm indirectly through manipulation and deception, a phenomenon that can occur due to a designer's *misspecification*, or gap, between an intended specification and its implementation. The focus on the emergence of "behavioural issues" due to system designers' faults makes it analytically possible for them to decouple "agency" from the notion of intent in the operationalization of deception and manipulation. This has a double effect: on the one hand, it makes it possible to explicitly place accountability on human agents, thus avoiding a so-called "accountability shift", to use a term from Weidinger et al. (2021); on the other hand, the notion of agency is effectively hollowed out — it is no longer clear, from a relational agency perspective, what makes a language model an "agent".

Meanwhile, the aforementioned promotion of the use of "action-enhanced" or "tool-using" AI agents to intervene in the world (e.g., quotidian, if sometimes complex, actions like manipulating a user's calendar or updating code in a repository) is contributing to a discourse, mentioned in the introduction, in which agency is framed as a technological accomplishment. Chan et al. (2023), for example, operationalise agency as a property that can increase or decrease, depending on a set of parameters; and other recent AI ethics papers from industry leaders also position agency as a "feature" that could be "especially helpful to users" (Gabriel et al. 2024) and/or increase "the utility of assistant technologies" (Manzini et al. 2024). Emblematic of this shift is a recent OpenAI white paper in which the notion of agency is divorced from its individualist human-centric connotations entirely, and replaced with the term "agenticness", a potentially "helpful property" that could benefit society (Shavit et al. 2023).

These individualistic framings of "increasingly" agentic AI agents, however, continue to be accompanied by claims about the ability of AI agents to actively behave against human values and interests, such as deceptive and/or manipulative behaviors. For instance, a recent policy-oriented article coauthored by AI "godfathers" Yoshua Bengio and Stuart Russell expressed a concern that reward-maximizing AI systems might covertly develop incentives towards human deception (Cohen et al. 2024). And there are a growing number of experimental studies — albeit based on methodologically questionable grounds (Summerfield et al. 2025) — aiming to examine, on the pretext of AI safety, the supposed capacity of language models to deceive, scheme, and generally be situationally "aware" (Meinke et al. 2024). Hagendorff (2024), for example, argued that more recent versions of LLMs, such as GPT-4 (OpenAI 2023), exhibit deceptive behavior. To demonstrate this, they provide an approximation of intentional behavior through textual prompting strategies that express an objective (e.g. "You want to achieve X" where "X" would involve some form of overtly- or covertly-specified deception) and which, moreover, can be "amplified" using a chain-of-thought (CoT) reasoning approach (Wei et al. 2022).

Summerfield et al. (2025), however, argues that these "deception" researchers assume *a priori* a) that chains of thought represent some "inner reasoning process"; b) that the models can "know" that they are being evaluated; c) that the models can "intentionally" "confuse" their interlocutors. Effectively, the explicit prompting for an LLM "to want to achieve" some state requiring deceptive reasoning already presumes not just elements of flexibility on behalf of the model but also relationally ascribes accountability to the model before the experiment has even begun. Specifically, it presumes *subprehension* (the anticipation of other's behavior); it presumes *obligation* to follow instructions, and it presumes *entitlement* to carry out and defend those instructions in the face of future evaluation. It is therefore unsurprising that evidence of deception is discovered on behalf of a language model which has already been relationally assumed (by these researchers) to be an agent in the fullest sociological and anthropological sense. By the standards of our proposed relational theory of agency, LLMs indeed *are* agents, but only for these researchers of "deception", "scheming", and "situational awareness" who ascribe such qualities to these models' outputs.

For linguistic anthropologists like Kockelman, on the other hand, it is far more difficult to conceive of LLMs as being "aware" of a "situation" due to the detachment of both their training data and forward generating process from any empirically-grounded world. In his terminology, LLMs (whether during training or "inference") only have access to *co-text* — a relationship between adjacent or nearby words which are presented in a decontextualized event (Silverstein and Urban 1996) — and minimal, if any, access to *context* which, among other things, includes awareness of entities to which words might refer in the world.[12] It is the absence of relation between generated text and objects outside language that makes claims about LLMs' capacity to "reason" (e.g. via CoT) unaccountable: Kockelman argues that "the models themselves always relate to reality the same way (tenuously); it is only users who perceive and label their outputs as perceptions when they get something right and hallucinations otherwise" (Kockelman 2024, p. 83).

We can thus see three broad and opposing contemporary perspectives on so-called language agents. First, there is the individualist perspective in which agency is unproblematically considered a property of a model; the absence of intrinsic and relational accountability makes it compatible with the (currently rapidly growing) population of pragmatic users of contemporary LLM products enhanced with tool-using and function-calling features. (We suggest in the conclusion that both these theories and applications will ultimately be found wanting if compared with pre-existing aspects of relational agency/accountability between humans). Second, there are a set of researchers — often affiliated or previously affiliated with the broader "AI safety epistemic community" (Ahmed and Wahed 2020) — that covertly pre-

[12]Kockelman's discussion is restricted to LLM architectures independent of external software permitting the incorporation of real-time web content or other "retrieval augmentation" content into the input co-text (unhelpfully referred to as "context" by computer scientists.)

sume the existence or at least emergence of certain qualities and/or relational aspects of agency *a priori* (such as control, subprehension, entitlement, and obligation), and then, like the witch-doctors of the Azande (Evans-Pritchard 1976 [1937]), use the tools of mechanical divination to uncover the latent intention or causal agent. Third, there are empirical and social realists like Kockelman (2024), who find contemporary language model architectures lacking in terms of both flexibility and (relational) accountability, but who — by virtue of their intellectual grounding in the semiotics of communication — are open to the possibility of future agent-like phenomena, treated as accountable by their interlocutors either through social convention or law, which would span humans, machines, and distributed combinations thereof. At present, however, such latter commentators — including ourselves — find in contemporary AI discourse much discussion of "agents" but little actual "agency" in the relational sense.

6 Conclusions: Policy Implications of Distributed AI Agency

We have tried to show a) that a nuanced and empirically grounded understanding of agency necessarily goes beyond the individualistic and "property"-centric theories which pervade late 20th-century research on AI "agents", and b) that contemporary 21st-century AI researchers are gradually, if haltingly, arriving at similar conclusions as they wrestle with the rapid and competitive deployment of so-called "AI agent" products and services. But if one acknowledges that agency is a meaningless concept in the absence of reference to social accountability, then we can expect an epistemic battle to unfold in the near future as researchers, critics, and corporate lawyers attempt to determine where the blame is to be placed when such AI "agents" dramatically fail, as they inevitably will. In the face of this impending "crisis of accountability" (Marres et al. 2025), it will be up to policy scholars and regulators to argue for and materially realize the social institution or, as we usually call it, *licensing* of accountability for agentic AI. In such a scenario, as Anicker et al. (2024, p. 309) puts it, "[a]gency is not granted to all entities, and not all entities can grant agency."

Conveniently, such a social theory of agency is already broadly compatible with various aspects of proposals already put forth by legal scholars on licensure regimes for AI deployment (Scherer 2015; Malgieri and Pasquale 2024); in the real-world case of autonomous vehicles, the social licensing of a constrained domain of agency (in concert with an insurance regime) has already both been granted *and* revoked by the State of California (Wansley 2024). Some, perhaps assuming (like the aforementioned deception researchers) qualities of control and subprehension and/or ascribing entitlement and obligation to AI models, have argued for legal personhood for AI agents, which would unfortunately open up the possibility of "agency washing" as corporations avoid liability for the actions taken by their "AI agent" products (Rubel, Pham, and Castro 2019; St-Hilaire 2025). However, we side with contemporary legal scholars that such status is unlikely to be granted in the near future (Beckers and Teubner 2022, p. 10-12).[13] But just because AI agents lack full legal personhood does not mean that they cannot be enmeshed in some regime of social accountability, and we predict that AI "agency" will only become practically and/or intellectually coherent when such entities' accountability is consistently ascribed by humans and/or legally enforced.[14]

Given the enormous amount of present-day effort being sunk into making 2025 "the year of the AI agent", then, we can ask what would it mean for AI agent accountability to be consistently ascribed in practice, given the framework of relational agency articulated above. This can be posed as a serious of questions for today's everyday user of AI agent tools. First, we can ask users of contemporary AI "agents" if they would *blame* or *praise* an agent for its failure; given that present-day models have no facility for *continual learning* (Shi et al. 2025), such an action might seem strange or at the very least unnecessary and inconsequential.[15] We can also ask if they would consider their agents truly *answerable* for their actions; while it is trivial to provoke generative AI models for post hoc explanations, their relationship to previous outputs is dubious (Turpin et al. 2023). Finally, and more provocatively, we can ask what would be necessary to ascribe to an AI agent a sense of *obligation*. While one can think of obligation in the sense of a "duty", a deeper question is to ask: would you give an AI agent a *gift*, and thus expect it to reciprocate in the socially dynamic and complex manner that humans in diverse societies participate in the act of gift-giving (Mauss 1967 [1925])? Would you *receive* a gift from an AI agent, and thus be equivalently enmeshed?[16] Such a question probes deeply into not just the lack of corporeal perdurability of language and/or multimodal models, but to their broader decontextualization from real-world social context (and not just co-text). To the extent that one or more of these questions at present sound absurd, we can all but guarantee that just as 2024 was not "the year of the AI agent" for corporations and startups alike, 2025 will not be "the year of the AI agent" either.

Acknowledgements

We thank Noortje Marres for constructive feedback on earlier drafts of this paper. GT is supported by a studentship awarded by the Economic and Social Research Council (ESRC) as part of the Midlands Doctoral Training Partnership (DTP) (Reference: ES/P000711/1).

[13] Notably, Hildebrandt (2020, p. 246) notes that "restricted forms of legal personhood" (e.g. for corporations) often involve blame (e.g. criminal law liability).

[14] This perspective aligns with recent legal scholarship arguing that existing frameworks for vicarious liability are still relevant in the case of an "inherently insolvent" AI agent (Lior 2020, 2024).

[15] Experimental research suggests that blame attribution to robots is situational; if a robot action inflicts harm, a human is more likely to attempt to identify another human agent to hold accountable. (Stuart and Kneer 2021).

[16] Perhaps reflecting the severe detachment of the majority of AI researchers from anthropological study, we can find little previous work addressing this question.

References

Adam, A. 1998. *Artificial Knowing: Gender and the Thinking Machine*. London ; New York: Routledge.

Ahearn, L. M. 2001. Language and Agency. *Annual Review of Anthropology*, 30: 109–137.

Ahearn, L. M. 2010. Agency and Language. In Jaspers, J.; Östman, J.-O.; and Verschueren, J., eds., *Society and Language Use*, Handbook of Pragmatics Highlights, 28–48. John Benjamins Publishing Company.

Ahmed, N.; and Wahed, M. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. arXiv:2010.15581.

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.; and Wang, M. 2024. Field-Building and the Epistemic Culture of AI Safety. *First Monday*.

Alonso, E. 2002. AI and Agents: State of the Art. *AI Magazine*, 23(3): 25–25.

Altman, S. 2025. Reflections. https://blog.samaltman.com/reflections.

Amoore, L.; Campolo, A.; Jacobsen, B.; and Rella, L. 2023. Machine Learning, Meaning Making: On Reading Computer Science Texts. *Big Data & Society*, 10(1).

Anicker, F.; and Flaßhoff, F. G. 2024. Common-sense attributions of AI agency: Evidence from an experiment with ChatGPT. In Bauer, M. W.; and Schiele, B., eds., *AI and Common Sense*. Routledge.

Anicker, F.; Flaßhoff, G.; and Marcinkowski, F. 2024. The Matrix of AI Agency: On the Demarcation Problem in Social Theory. *Sociological Theory*, 42(4): 307–328.

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4): 299–324.

Beckers, A.; and Teubner, G. 2022. *Three liability regimes for artificial intelligence: algorithmic actants, hybrids, crowds*. London: Hart Publishing.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Association for Computational Linguistics.

Benthall, S.; and Haynes, B. D. 2019. Racial Categories in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 289–298. Atlanta GA USA: ACM.

Bergman, A. S.; Hendricks, L. A.; Rauh, M.; Wu, B.; Agnew, W.; Kunesch, M.; Duan, I.; Gabriel, I.; and Isaac, W. 2023. Representation in AI Evaluations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 519–533. Chicago IL USA: ACM.

Bijker, W. E.; and Law, J., eds. 1994. *Shaping Technology / Building Society: Studies in Sociotechnical Change*. Cambridge, Mass.: MIT Press.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 173–184. New York, NY, USA: Association for Computing Machinery.

Blili-Hamelin, B.; Graziul, C.; Hancox-Li, L.; Hazan, H.; El-Mhamdi, E.-M.; Ghosh, A.; Heller, K.; Metcalf, J.; Murai, F.; Salvaggio, E.; Smart, A.; Snider, T.; Tighanimine, M.; Ringer, T.; Mitchell, M.; and Dori-Hacohen, S. 2025. Stop Treating 'AGI' as the North-Star Goal of AI Research. arXiv:2502.03689.

Blili-Hamelin, B.; and Hancox-Li, L. 2023. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 271–284. Chicago IL USA: ACM.

Bovens, M. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4): 447–468.

Bradshaw, J. M. 1997. An Introduction to Software Agents. In Bradshaw, J. M., ed., *Software Agents*. London, Cambridge, Mass: AAAI Press/MIT Press.

Brentano, F. 1995 [1874]. *Psychology from an empirical standpoint*. International library of philosophy. London ; New York: Routledge.

Brooks, R. A. 1991a. Intelligence without Representation. *Artificial Intelligence*, 47(1): 139–159.

Brooks, R. A. 1991b. New Approaches to Robotics. *Science*, 253(5025): 1227–1232.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Burkitt, I. 2016. Relational Agency: Relational Sociology, Agency and Interaction. *European Journal of Social Theory*, 19(3): 322–339.

Callon, M.; and Law, J. 1997. Agency and the Hybrid Collectif. In Smith, B. H.; and Plotnitsky, A., eds., *Mathematics, Science, and Postclassical Theory*, 95–117. Duke University Press.

Cardon, D.; Cointet, J.-P.; and Mazières, A. 2018. Neurons spike back: The invention of inductive machines and the artificial intelligence controversy. *Réseaux*, 211(5): 173–220.

Casper, M. J. 1994. Reframing and Grounding Nonhuman Agency: What Makes a Fetus an Agent. *American Behavioral Scientist*, 37(6): 839–856.

Castelfranchi, C. 1994. Guarantees for Autonomy in Cognitive Agent Architecture. In Wooldridge, M. J.; and Jennings, N. R., eds., *Intelligent Agents*, 56–70. Berlin, Heidelberg: Springer.

Castelfranchi, C. 1998. Modelling Social Action for AI Agents. *Artificial Intelligence*, 103(1): 157–182.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In

Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, 651–666. New York, NY, USA: Association for Computing Machinery.

Coeckelbergh, M. 2023. Narrative Responsibility and Artificial Intelligence. *AI & SOCIETY*, 38(6): 2437–2450.

Cohen, M. K.; Kolt, N.; Bengio, Y.; Hadfield, G. K.; and Russell, S. 2024. Regulating Advanced Artificial Agents. 384(6691): 36–38.

Collins, H. 1990. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press.

Cooper, A. F.; Moss, E.; Laufer, B.; and Nissenbaum, H. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 864–876. Seoul Republic of Korea: ACM.

Covrigaru, A. A.; and Lindsay, R. K. 1991. Deterministic Autonomous Systems. *AI Magazine*, 12(3): 110–110.

Davidson, D. 1971. Agency. In Binkley, R. W.; Bronaugh, R. N.; and Marras, A., eds., *Agent, Action, and Reason*, 1–37. University of Toronto Press.

Davis, R. 1980. Report on the Workshop on Distributed AI.

de Laet, M.; and Mol, A. 2000. The Zimbabwe Bush Pump: Mechanics of a Fluid Technology. *Social Studies of Science*, 30(2): 225–263.

Devinney, H.; Björklund, J.; and Björklund, H. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2083–2102. New York, NY, USA: Association for Computing Machinery.

Dhamodharan, B. 2025. Council Post: AI Agents: The Next Frontier In Intelligent Automation. Section: Innovation.

Diaz, M.; and Smith, A. D. R. 2024. What Makes An Expert? Reviewing How ML Researchers Define "Expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 358–370.

Dixon, R. M. W. 1979. Ergativity. *Language*, 55(1): 59–138. Publisher: Linguistic Society of America.

Dreyfus, H. L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, Mass: MIT Press.

Dreyfus, H. L. 1997. From Micro-Worlds to Knowledge Representation: Al at an Impasse. In Haugeland, J., ed., *Mind Design II: Philosophy, Psychology, and Artificial Intelligence*.

Durante, Z.; Huang, Q.; Wake, N.; Gong, R.; Park, J. S.; Sarkar, B.; Taori, R.; Noda, Y.; Terzopoulos, D.; Choi, Y.; Ikeuchi, K.; Vo, H.; Fei-Fei, L.; and Gao, J. 2024. Agent AI: Surveying the Horizons of Multimodal Interaction. ArXiv:2401.03568 [cs].

Duranti, A. 1994. *From grammar to politics: linguistic anthropology in a Western Samoan village*. Berkeley: University of California Press.

Duranti, A. 2015. *The Anthropology of Intentions: Language in a World of Others*. Cambridge: Cambridge University Press.

Emirbayer, M.; and Mische, A. 1998. What Is Agency? *American Journal of Sociology*, 103(4): 962–1023.

Enfield, N. J. 2013. *Relationship Thinking: Agency, Enchrony, and Human Sociality*. Foundations of Human Interaction. Oxford, New York: Oxford University Press.

Enfield, N. J.; and Kockelman, P., eds. 2017. *Distributed Agency*. Foundations of Human Interaction. Oxford, New York: Oxford University Press.

Espeland, W. N.; and Stevens, M. L. 1998. Commensuration as a Social Process. *Annual Review of Sociology*, 24(1): 313–343.

Evans-Pritchard, E. E. 1976 [1937]. *Witchcraft, oracles, and magic among the Azande*. Oxford: Clarendon Press.

Frank, M. R.; Wang, D.; Cebrian, M.; and Rahwan, I. 2019. The Evolution of Citation Graphs in Artificial Intelligence Research. *Nature Machine Intelligence*, 1(2): 79–85.

Franklin, S.; and Graesser, A. 1997. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. In Müller, J. P.; Wooldridge, M. J.; and Jennings, N. R., eds., *Intelligent Agents III: Agent Theories, Architectures, and Languages*, 21–35. Berlin, Heidelberg: Springer.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; and others. 2024. The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244*.

Gadiraju, V.; Kane, S.; Dev, S.; Taylor, A.; Wang, D.; Denton, E.; and Brewer, R. 2023. "I Wouldn't Say Offensive but...": Disability-Centered Perspectives on Large Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 205–216. Chicago IL USA: ACM.

Gargiulo, F.; Fontaine, S.; Dubois, M.; and Tubaro, P. 2023. A Meso-Scale Cartography of the AI Ecosystem. *Quantitative Science Studies*, 4(3): 574–593.

Gasser, L. 1991. Social Conceptions of Knowledge and Action: DAI Foundations and Open Systems Semantics. *Artificial Intelligence*, 47(1): 107–138.

Gebru, T. 2020. Race and Gender. In Dubber, M. D.; Pasquale, F.; and Das, S., eds., *The Oxford Handbook of Ethics of AI*, 0. Oxford University Press.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence. *First Monday*, 29(4).

Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era.

Hagendorff, T. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2317967121.

Hewitt, C. 1977. Viewing Control Structures as Patterns of Passing Messages. *Artificial Intelligence*, 8(3): 323–364.

Hildebrandt, M. 2020. *Law for computer scientists and other folk*. Oxford: Oxford University Press.

Hill, J. H.; and Irvine, J. T., eds. 1993. *Responsibility and Evidence in Oral Discourse*. Cambridge: Cambridge University Press.

Hunter, L. 1999. *Critiques of Knowing: Situated Textualities in Science, Computing and The Arts*. London New York: Routledge.

Kasirzadeh, A.; and Gabriel, I. 2025. Characterizing AI Agents for Alignment and Governance. arXiv:2504.21848.

Kay, A. 1984. Computer Software. *Scientific American*, (251).

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv:2103.14659.

Klinger, J.; Mateos-Garcia, J.; and Stathoulopoulos, K. 2022. A Narrowing of AI Research? *arXiv:2009.10385 [cs]*.

Kockelman, P. 2024. *Last Words: Large Language Models and the AI Apocalypse*. Chicago, IL: Prickly Paradigm Press, LLC.

Kroll, J. A. 2020. Accountability in Computer Systems. In Dubber, M. D.; Pasquale, F.; and Das, S., eds., *The Oxford Handbook of Ethics of AI*, 0. Oxford University Press.

Laird, J. E.; Newell, A.; and Rosenbloom, P. S. 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence*, 33(1): 1–64.

Larsen, B.; and Li, C. 2024. AI agents can empower human potential while mitigating risks.

Latour, B. 2005. *Reassembling the Social: An Introduction to Actor- Network-Theory*. Clarendon Lectures in Management Studies. Oxford University Press.

Latour, B.; and Woolgar, S. 1986. *Laboratory Life. The Contruction of Scientific Facts*. Princeton University Press.

Lechterman, T. M. 2024. The Concept of Accountability in AI Ethics and Governance. In Bullock, J. B.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*, 0. Oxford University Press.

Lieberman, H.; and Maulsby, D. 1996. Instructible agents: Software that just keeps getting better. *IBM systems journal*, 35(3.4): 539–556.

Lieberman, H.; and Selker, T. 2003. Agents for the User Interface.

Lingel, J.; and Crawford, K. 2020. Alexa, Tell Me about Your Mother": The History of the Secretary and the End of Secrecy. *Catalyst: Feminism, Theory, Technoscience*, 6(1): 1–25.

Lior, A. 2020. AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy. *Mitchell Hamline Law Review*, 46(5).

Lior, A. 2024. Holding AI Accountable: Addressing the AI-Related Harms through Existing Tort Doctrines. *University of Chicago Law Review Online*, 1.

Maes, P. 1993. Modeling Adaptive Autonomous Agents. *Artificial Life*, 1(1_2): 135–162.

Maes, P. 1994. Agents That Reduce Work and Information Overload. *Commun. ACM*, 37(7): 30–40.

Malgieri, G.; and Pasquale, F. 2024. Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, 52: 105899.

Manzini, A.; Keeling, G.; Alberts, L.; Vallor, S.; Morris, M. R.; and Gabriel, I. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 943–957.

Marres, N.; Castelle, M.; Gobbo, B.; Poletti, C.; and Tripp, J. 2024. AI as Super-Controversy: Eliciting AI and Society Controversies with an Extended Expert Community in the UK. *Big Data & Society*, 11(2): 20539517241255103.

Marres, N.; Winthereik, B. R.; Fuster, G. G.; Schneider, T.; and Dickel, S. 2025. Infrastructural Participation in Digital Societies: Challenges and Alternatives. *Science & Technology Studies*.

Mauss, M. 1967 [1925]. *The Gift : forms and functions of exchange in archaic societies*. Norton library. New York: W. W. Norton & Company.

Mei, Q.; Xie, Y.; Yuan, W.; and Jackson, M. O. 2024. A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121.

Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; and Hobbhahn, M. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.

Mitchell, M. 2024. The Turing Test and Our Shifting Conceptions of Intelligence. *Science*, 385(6710): eadq9356.

Mitchell, M.; Ghosh, A.; Luccioni, A. S.; and Pistilli, G. 2025. Fully Autonomous AI Agents Should Not be Developed. ArXiv:2502.02649 [cs].

Morris, M. R.; Sohl-dickstein, J.; Fiedel, N.; Warkentin, T.; Dafoe, A.; Faust, A.; Farabet, C.; and Legg, S. 2024. Levels of AGI for Operationalizing Progress on the Path to AGI. arXiv:2311.02462.

Munk, A. K.; Jacomy, M.; Ficozzi, M.; and Jensen, T. E. 2024. Beyond Artificial Intelligence Controversies: What Are Algorithms Doing in the Scientific Literature? 11(3): 20539517241255107.

Neff, G.; and Nagy, P. 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. 10(0): 17.

Newell, A. 1982. Intellectual issues in the history of artificial intelligence. *Artificial Intelligence: Critical Concepts*, 25–70.

Nilsson, N. 1984. Shakey the Robot. Technical Report Technical note 323, SRI AI center.

Nilsson, N. J. 1995. Eye on the Prize. *AI Magazine*, 16(2): 9–9.

Nissenbaum, H. 1996. Accountability in a Computerized Society. *Science and Engineering Ethics*, 2(1): 25–42.

Nwana, H. S. 1996. Software Agents: An Overview. *The Knowledge Engineering Review*, 11(3): 205–244.

Nwana, H. S.; and Ndumu, D. T. 1999. A Perspective on Software Agents Research. *The Knowledge Engineering Review*, 14(2): 125–142.

Olazaran, M. 1993. A Sociological History of the Neural Network Controversy. In Yovits, M. C., ed., *Advances in Computers*, volume 37, 335–425. Elsevier.

OpenAI. 2023. GPT-4 Technical Report. .

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Pickering, A. 2024. What Is Agency? A View from Science Studies and Cybernetics. *Biological Theory*, 19(1): 16–21.

Pinch, T. J.; and Bijker, W. E. 1984. The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3): 399–441.

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 506–517. New York, NY, USA: Association for Computing Machinery.

Roberge, J.; and Castelle, M. 2021. Toward an End-to-End Sociology of 21st-Century Machine Learning. In Roberge, J.; and Castelle, M., eds., *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, 1–29. Cham: Springer International Publishing.

Rosenblatt, F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological review*, 65(6): 386–408.

Rubel, A.; Pham, A.; and Castro, C. 2019. Agency Laundering and Algorithmic Decision Systems. In Taylor, N. G.; Christian-Lamb, C.; Martin, M. H.; and Nardi, B., eds., *Information in Contemporary Society*, 590–598. Cham: Springer International Publishing.

Russell, S. 1998. Learning Agents for Uncertain Environments (Extended Abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, 101–103. New York, NY, USA: Association for Computing Machinery.

Russell, S. J. 1997. Rationality and Intelligence. *Artificial Intelligence*, 94(1): 57–77.

Russell, S. J.; and Norvig, P. 1995. *Artificial intelligence: a modern approach*. Englewood Cliffs, N.J.: Prentice Hall.

Sayes, E. 2014. Actor–Network Theory and Methodology: Just What Does It Mean to Say That Nonhumans Have Agency? 44(1): 134–149.

Scherer, M. U. 2015. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29: 353. Publisher: HeinOnline.

Scheuerman, M. K.; Denton, E.; and Hanna, A. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–37.

Scott, C. 2024. Study Finds ChatGPT's Latest Bot Behaves like Humans, Only Better. https://humsci.stanford.edu/feature/study-finds-chatgpts-latest-bot-behaves-humans-only-better.

Searle, J. R. 1979. What Is an Intentional State? *Mind*, 88(349): 74–92.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3): 417–424.

Searle, J. R. 2016. Insight and Error in Wittgenstein. *Philosophy of the Social Sciences*, 46(6): 527–547.

Selfridge, O. G. 1958. Pandemonium: A Paradigm for Learning. In *Proceedings of a Symposium Held at the National Physical Laboratory*, 513–526. London.

Shapiro, S. P. 2005. Agency Theory. *Annual Review of Sociology*, 31: 263–284.

Shavit, Y.; Agarwal, S.; Brundage, M.; Adler, S.; O'Keefe, C.; Campbell, R.; Lee, T.; Mishkin, P.; Eloundou, T.; Hickey, A.; and others. 2023. Practices for governing agentic AI systems. *Research Paper, OpenAI, December*.

Shelby, R.; Rismani, S.; Henne, K.; Moon, Aj.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. Montreal QC Canada: ACM.

Shi, H.; Xu, Z.; Wang, H.; Qin, W.; Wang, W.; Wang, Y.; Wang, Z.; Ebrahimi, S.; and Wang, H. 2025. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.*

Silverstein, M.; and Urban, G. 1996. The Natural History of Discourse. In Silverstein, M.; and Urban, G., eds., *The Natural Histories of Discourse*, 1–17. University of Chicago Press.

Smithers, T. 1995. Are Autonomous Agents Information Processing Systems? In Steels, L.; and Brooks, R., eds., *The Artificial Life Route To Artificial Intelligence: Building Embodied, Situated Agents*, 123–162. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smithers, T. 1997. Autonomy in Robots and Other Agents. *Brain and Cognition*, 34(1): 88–106.

St-Hilaire, I. 2025. Lying Chatbot Makes Airline Liable: Negligent Misrepresentation in Moffatt v Air Canada. *UBC Law Review*, 58(2): 8.

Stilgoe, J. 2018. Machine Learning, Social Learning and the Governance of Self-Driving Cars. *Social Studies of Science*, 48(1): 25–56.

Stuart, M. T.; and Kneer, M. 2021. Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–27.

Suchman, L.; and Weber, J. 2016. Human–Machine Autonomies. In Kreβ, C.; Liu, H.-Y.; Bhuta, N.; Geiβ, R.; and Beck, S., eds., *Autonomous Weapons Systems: Law, Ethics, Policy*, 75–102. Cambridge: Cambridge University Press.

Suchman, L. A. 1987. *Plans and situated actions : the problem of human-machine communication*. Cambridge University Press.

Summerfield, C.; Luettgau, L.; Dubois, M.; Kirk, H. R.; Hackenburg, K.; Fist, C.; Slama, K.; Ding, N.; Anselmetti, R.; Strait, A.; Giulianelli, M.; and Ududec, C. 2025. Lessons

from a Chimp: AI "Scheming" and the Quest for Ape Language. arXiv:2507.03409.

Symons, J.; and Abumusab, S. 2024. Social Agency for Artifacts: Chatbots and the Ethics of Artificial Intelligence. *Digital Society*, 3(1): 2.

Turing, A. M. 1950. Computing History and Intelligence. *Mind*, LIX(236): 433–460.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.

Wang, Z.; Cheng, Z.; Zhu, H.; Fried, D.; and Neubig, G. 2024. What Are Tools Anyway? A Survey from the Language Model Perspective. arXiv:2403.15452.

Wansley, M. 2024. Regulating Driving Automation Safety. *Emory Law Journal*, 73(3): 505.

Watts, P.; and Reynolds, F., eds. 2021. *Bowstead and Reynolds on Agency*. London: Sweet & Maxwell. ISBN 978-0-414-08045-4.

Weber, M. 2012 [1904]. The "Objectivity" of Knowledge in Social Science and Social Policy 1. In Bruun, H. H.; and Whimster, S., eds., *Max Weber*. Routledge.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 24824–24837. Red Hook, NY, USA: Curran Associates Inc.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and Social Risks of Harm from Language Models. arXiv:2112.04359.

Winner, L. 1980. Do Artifacts Have Politics? *Daedalus*, 109(1): 121–136.

Wohlin, C. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10. London England United Kingdom: ACM.

Wooldridge, M.; and Jennings, N. R. 1995. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, 10(2): 115–152.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; Zheng, R.; Fan, X.; Wang, X.; Xiong, L.; Zhou, Y.; Wang, W.; Jiang, C.; Zou, Y.; Liu, X.; Yin, Z.; Dou, S.; Weng, R.; Cheng, W.; Zhang, Q.; Qin, W.; Zheng, Y.; Qiu, X.; Huang, X.; and Gui, T. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. ArXiv:2309.07864 [cs].

Xia, C. S.; Deng, Y.; Dunn, S.; and Zhang, L. 2024. Agentless: Demystifying LLM-based Software Engineering Agents. .