Synthese (2022) 200:463
https://doi.org/10.1007/s11229-022-03902-9

ORIGINAL RESEARCH

Decentring the discoverer: how AI helps us rethink scientific discovery

Elinor Clark[1] · **Donal Khosrowi**[1]

Received: 15 October 2021 / Accepted: 22 September 2022 / Published online: 3 November 2022

Abstract

This paper investigates how intuitions about scientific discovery using artificial intelligence (AI) can be used to improve our understanding of scientific discovery more generally. Traditional accounts of discovery have been *agent-centred*: they place emphasis on identifying a specific agent who is responsible for conducting all, or at least the important part, of a discovery process. We argue that these accounts experience difficulties capturing scientific discovery involving AI and that similar issues arise for human discovery. We propose an alternative, *collective-centred* view as superior for understanding discovery, with and without AI. This view maintains that discovery is performed by a collective of agents and entities, each making contributions that differ in significance and character, and that attributing credit for discovery depends on various finer-grained properties of the contributions made. Detailing its conceptual resources, we argue that this view is considerably more compelling than its agent-centred alternative. Considering and responding to several theoretical and practical challenges, we point to concrete avenues for further developing the view we propose.

Keywords AI · Scientific discovery · Agent-centred view · Collective-centred view · AlphaFold

✉ Donal Khosrowi
donal.khosrowi@philos.uni-hannover.de

Elinor Clark
elinor.clark2@gmail.com

[1] Institute of Philosophy, Leibniz Universität Hannover, Hannover, Germany

1 Introduction

Researchers across the sciences increasingly explore the potential of artificial intelligence (AI)[1] systems to make or assist with scientific discoveries.[2] Existing attempts range from early programs such as BACON to 'rediscover' Kepler's third law, Coulomb's law and Ohm's law (Langley, 1977), to predicting protein structures with DeepMind's AlphaFold (Jumper et al., 2021; Senior et al., 2019); mining archival materials to understand the human impact of the Industrial Revolution (Ardanuy et al., 2019); and predicting future sea ice coverage by uncovering previously unrecognised relationships from Arctic sea ice data (Banerjee & Monteleoni, 2014).

In light of these advances, it can be tempting to think that AI systems are already making, or are at least along the path towards, full-fledged scientific discoveries (Bannigan et al., 2021; Häse et al., 2019; MacLeod et al., 2020). However, other authors are more sceptical and offer reasons for thinking AI systems cannot discover, pointing to limitations such as lack of domain-general knowledge or imaginative abilities (Halina, 2021; Stuart, 2019). We do not think that questions of whether and how AI systems (can) make discoveries are quite ready to be settled yet, however, and that this is in part due to important deficiencies of existing views of scientific discovery. Our main aim in this paper is to articulate these deficiencies and draw on case-based intuitions about discovery involving AI to develop a novel view that improves our understanding of scientific discovery, with and without AI.

Specifically, we consider an influential view of scientific discovery that continues to prominently shape discovery narratives: the *agent-centred view* (AC). AC seeks to identify a central agent (or small group of agents) responsible for a discovery and explains the success of discovery by reference to specific qualities and abilities of these agents. This view has been criticised extensively (e.g. Copeland, 2018; Zytkow, 1996), and we add to these criticisms by exploring how AC struggles with adequately capturing the role of AI systems in discovery, as well as discovery sans AI more generally.

With AC's shortcomings articulated, we propose an alternative, *collective-centred view* (CC), which insists that focusing on individual agents as discoverers is misguided and obscures important features of scientific discovery. Instead, CC maintains that discovery is performed by a collective of agents and entities, each making contributions that differ in significance and character, and where membership in the collective and individual stakes in a discovery depend on various finer-grained properties of the contributions made.

[1] Throughout the paper, we use the label 'AI' to focus on current state-of-the-art machine learning-based systems, such as DeepMind's AlphaFold (Jumper et al., 2021), as they are used in scientific contexts.

[2] Various definitions of 'scientific discovery' have been proposed (Brannigan, 1981; Hanson, 1960; Magnani, 2000; Reichenbach, 1938; Schiller, 1917; Whewell, 1996 [1840]). We are not concerned here with clarifying what counts as a discovery, or how scientific communities come to answer this question, but will assume that the cases discussed here are indeed cases of scientific discovery and focus on how different agents and entities play a role in these. We also gloss over the general distinction between *serendipitous* and *purposeful* discovery, i.e. discovering things one is not actively searching for (Copeland, 2018 p. 697) versus strategically pursuing a specific discovery (e.g. of the Higgs boson). Our discussion focuses on the latter, mainly since existing cases of AI-based discovery often take such forms, but our arguments apply to serendipitous discovery too.

Beyond elucidating the role that AI plays in discovery, we argue that CC also provides a more plausible view of scientific discovery more generally, and highlight how it can promote ameliorative efforts to (re-)characterise discovery episodes in ways that are not only descriptively more adequate but also allocate credit more justly. CC hence follows a rich tradition of emphasising the intrinsically social nature of scientific enterprises, including discovery (Brannigan, 1981; Kuhn, 1970; Longino, 1990; Zytkow, 1996). It also builds on, and reinforces, efforts to expose the historically unrepresentative nature of discovery narratives that pick out a central discoverer, usually a 'guy in a lab', and neglect contributions of other actors (Copeland, 2018; Qureshi et al., 2021; Schiebinger, 1987), as well as arguments emphasising and analysing the historic and present-day injustices experienced by female and minority groups in the context of knowledge production and beyond (Fricker, 2007; Mills, 2007). At the same time, our work extends significantly beyond these contributions by developing CC as a concrete alternative to AC. Although important challenges remain in applying CC in practice, we outline how it can stimulate ongoing efforts to capture the roles AI may play in discovery, and improve our general philosophical understanding of scientific discovery.

We proceed as follows. Section 2 characterises AC and highlights its central tenets. Section 3 considers three ways of applying AC to AI discovery, explains how they are unsuccessful, and considers additional difficulties experienced by AC in capturing human discovery. Section 4 presents CC, details its resources for identifying and appraising contributions to discovery, and explores how it helps us better understand AI discovery. Section 5 considers some theoretical and practical challenges for CC, offers suggestions for how they could be resolved, and gestures to ways of developing CC further. Section 6 concludes.

2 Agent-centred views

Alexander Fleming's discovery of penicillin in 1928 (Fleming, 1929) is considered one of the most significant scientific discoveries of the twentieth century. Famously, Fleming carelessly left out a petri dish of staphylococcus in the lab when he went on holiday. On his return, he noticed a mould had developed and, when examining the culture, realised the mould had prevented the staphylococci from growing. Fleming's realisation is often presented as a crucial discovery event that paved the way for developing the most widely used antibiotic in the world and helped save millions of lives.

In understanding discovery events like these, influential philosophical accounts of scientific discovery have historically focused on identifying a specific agent who is the 'discoverer' and describing this person's experience of discovery (Klahr & Simon, 1999; Schaffer, 1986). For instance, Whewell (1996 [1840]) emphasises the 'eureka moment', a sudden flash of insight and inspiration which sparks a discovery. In Whewell's words, discovery centrally involves "[…] some happy thought, of which we cannot trace the origin; some fortunate cast of intellect, rising above all rules." (Whewell, 1996 [1840], p. 186). Other cases abound that fit this schema: take Ignaz Semmelweis' discovery of the cause of puerperal fever (Best & Neuhauser, 2004;

Shorter, 1984) or Yoshida et al.'s discovery of plastic-eating bacteria (2016). Importantly, a happy thought will only arise in the mind of an agent who is able and willing to see it: the discoverer.

A more recent view is offered by Stuart (2019), who articulates conditions for AI discovery in social science and identifies the following elements of a discovery event: "(1) an agent (who discovers), (2) an object of discovery (that which is discovered), (3) a trigger event (that which prompts the discovery) and (4) an act of discovery (the agent's interpretation of the object, prompted by the trigger event)" (2019, p. 4).

These views share a distinctive emphasis on identifying a discovering agent who is central to a discovery, responsible for it, and to be credited with it. Call this the *agent-centred view* (henceforth AC). AC asserts that scientific discoveries usually conform to the following schema:

(1) *There is a single scientist or small research team we can pick out as the* **relevant discovering agent**.
(2) *This discovering agent* **conducted all, or at least the important part, of the discovery process**.
(3) *The discovering agent has particular qualities/abilities which* **play a significant role** *in the discovery*.

Being able to identify a discovering agent is important: it allows us to incentivise and credit those responsible for a discovery, as well as hold them to account if the need arises (e.g. Urbina et al., 2022).[3] AC helps with this: it identifies agents who conduct the process that yields a discovery and explains the success of this process in terms of agents' qualities and abilities that are essential to a discovery.[4] This puts clear constraints on who counts as a discoverer: for one, you are either in or out; being a discoverer is a dichotomous, not a gradual affair. Moreover, the bar to get in is high: you need to deploy particular qualities and abilities to significant effect, otherwise you are part of the background at best.

Unsurprisingly, this type of view has attracted significant suspicion and criticism from sociologists and philosophers of science. For instance, Schaffer (1986) points out that philosophers have used idealised versions of iconic discoveries, such as of Oxygen, Uranus, and photosynthesis, to shape their philosophical accounts, taking discoveries "[…] rather unproblematically, as single events of individual mental labour whose analysis requires the examination of logical or psychological manoeuvres." (Schaffer, 1986, p. 388). However, as he goes on to argue, it is rarely the case that discovery is a single-authored event, even for those discoveries seemingly archetypal of AC. More

[3] While holding discoverers to account for negative consequences of their discoveries is important, we focus only on the positive side of credit allocation.

[4] It is sometimes not fully clear what abilities, in particular, AC takes to be central to discovery. For instance, Stuart (2019) emphasises that producing novel, significant interpretations of events may require specific imaginative abilities. Whewell is less precise, emphasising a "special process on the mind" (1849, p. 40) and an ability to bring together a set of facts to afford a novel interpretation. Copeland (2018) offers an alternative, skills-based account, where even serendipitous discoveries can be strategically promoted through "the exercise of specific types of perception and attention that can in turn be cultivated" (2018, p. 715). Our arguments apply across different ways of casting what abilities and qualities AC precisely takes to be central to discovery, be they cognitive, mental abilities, including abilities to observe, connect observations, interpret them in novel ways, and to recognise their significance, or broader virtues such as curiosity and inquisitiveness.

recently, Copeland (2018) emphasises that AC "[…] tends to obscure the epistemology of discovery and to impede discussion about the importance of diffusing epistemic credit for discovery among members of the contributing network." (2018, p. 695).

We agree with Copeland's and Schaffer's assessments and follow them and other authors (Hull, 1988; Kukla, 2012; Longino, 1990; Nersessian, 1992; Zytkow, 1996) in maintaining that AC is unsatisfactory, as it neglects that science, by and large, is a social enterprise that often involves various agents playing different roles, each of which are important for enabling progress, including by discovery.

Adding to these criticisms, our first aim is to elaborate how AC specifically struggles with capturing the role that AI systems can play in scientific discovery. We argue that AC is unable to plausibly identify a discoverer in AI-based discovery, and is conceptually unprepared to explore and address questions about the distribution of credit among discoverers, what roles are played by different agents and entities, and what constitutes significant contributions to discovery. Following this, and reinforcing existing concerns, we argue that similar problems arise in the case of human discovery, too, and that AC is hence not suitable for fully understanding modern, or indeed historical, discovery.

3 AI discovery: where is our agent?

The protein folding problem is a challenge that has plagued biologists for decades (Dill et al., 2008). Each protein is comprised of chains of amino acids and has an intricate three-dimensional structure. Understanding these structures is crucial for making advances in drug development and improving our understanding of disease more generally. However, due to the complexity of interactions between amino acids and the intractable variety of possible conformations, predicting how these chains will fold is extremely time-consuming and costly.

To make progress on this challenge, researchers at DeepMind have developed AlphaFold[5] to speed up the process of predicting full protein structures (Jumper et al., 2021; Senior et al., 2019). At its core, AlphaFold is a transformer-based deep learning architecture that predicts whole three-dimensional protein structures from amino acid sequence inputs. Beyond achieving impressive gains in predictive accuracy over rival approaches in the recent community-wide structure prediction competition CASP14, AlphaFold has since been used to predict close to one million three-dimensional protein structures, which have been made freely available to the scientific community (Varadi et al., 2022) and will shortly be followed by almost 100 million additional structures. In light of these achievements, there is now widespread enthusiasm among academic and industry researchers about the potential of AlphaFold and similar, future AI systems to make important contributions to advances in a wide range of areas, including drug and antibody discovery (see e.g. Callaway, 2022; Mullard, 2021). Yet, as AlphaFold begins to play a central role in identifying protein structures, should we also conclude that it discovers them and, if so, in what sense?

[5] We focus here on v2.0 (Jumper et al., 2021).

We see three ways of applying AC to AI-based discovery. The first is to take AI systems to be discovering agents. The second is to say that humans discover and AI systems are mere tools. A third option considers both humans and AI systems as discoverers.[6] We consider each option in turn, arguing that none of them is fully compelling.

3.1 AI discovers

Can we say that AlphaFold is discovering on its own on AC? According to our construal of AC, AlphaFold would have to conduct all, or at least the important part, of the discovery process, as well as relevantly deploy abilities related to understanding and interpretation that are essential to the discovery. There are several problems with this option.

First, by picking out AlphaFold as the discovering agent, AC would fail to adequately consider human contributions essential to the discoveries made. AI systems do not (yet) determine their own research questions and commentators frequently underestimate how much supervision, clever engineering, trial and error, and fine-tuning is needed to make AI systems function as intended (Hagendorff & Wezel, 2020; Morgenstern, 2001; Parisi et al., 2019; Pearl, 2018; Sap et al., 2020). So focusing on AlphaFold alone misses essential parts of the discovery process.

Second, even if we assume AlphaFold's contribution marks the most important part of the discovery process, and human efforts take a back seat, it is at least unclear, and perhaps even unlikely, that AI systems like AlphaFold possess the important qualities and abilities essential to discovery emphasised by accounts like Whewell's and Stuart's.

Take Whewell's account: there is no 'happy thought', no sudden flash of insight from which the discovery is sparked. AI systems do not currently possess even foundational experiential abilities that could encode positive feelings and inspiration in response to surprising observations. Perhaps this simply suggests we should abandon emotive inflections in our conceptions of discovery, as these would uninterestingly preclude AI systems from being discoverers. However, there are other qualities and abilities important for discovery which AI systems seem to lack, even if we take out the emotive aspects. While AI systems are capable of generating novel solutions to significant problems, they lack domain-general knowledge, values and purposes, and are currently unable to guide their own research (Halina, 2021, p. 323). Relatedly, following Stuart (2019), discovery plausibly involves the ability to interpret and recognise significance. But AlphaFold does not know it is predicting protein structures to help synthesise drugs and cure disease. For lack of such context, AlphaFold will also remain unable to make contributions beyond its design envelope, such as through identifying follow-on research questions of interest, recognising the limitations of its findings, or suggesting avenues for applying its findings in practice. Looking beyond AlphaFold, lacking such abilities clearly restricts AI systems' capacity to make scientific discoveries by themselves.

[6] We thank an anonymous reviewer for suggesting that we consider this option more fully.

This is not to suggest, of course, that future AI systems could not be more self-directed in the various ways required by ordinary expectations for standalone discoverers (Buckner, 2018; Rafati & Noelle, 2019). Even so, the abilities of AI discoverers are likely to still depend crucially on assistance by human collaborators for the foreseeable future, if only concerning the values needed to steer discovery towards what humans judge as significant.

What this suggests, perhaps, is that whatever contributions to discovery AI systems make, these are not usefully captured by criteria that emphasise discoverers' abilities to independently understand, interpret, recognise significance, and so on. But, as we argue later, there are reasons to think that not all significant contributions to discovery must take such forms, whether made by machines or humans.

3.2 Humans discover, AI is a tool

A second way of applying AC to AI discovery asserts that humans do the discovering and AI systems are merely *tools*, more akin to a microscope or a particle detector than a human discoverer.

Our main concern with this option is that AC does not say much on how a distinction between genuine discovery and mere tool-like contributions could be made. One option would be to insist that genuine discoverers are those who, beyond making other crucial contributions, effectively deploy their interpretive abilities to recognise significance. But this seems an implausible standard. For instance, in large-scale research projects (e.g. at CERN) it would seem odd to identify those lead researchers conveniently situated to synthesise conclusions from varied results as discoverers, while discounting the role played by other researchers furnishing those results to those of mere tools. Conversely, it is often not principal investigators but less senior researchers who simultaneously do tedious groundwork *and* furnish interpretations of results. So this option is doubly at odds with how credit is plausibly shared in large-scale research projects.

To help AC draw a distinction between genuine discovery and mere tool-like contributions, important considerations could include (1) how relevant and replaceable a contribution is, (2) how much time and effort it requires, and (3) how much human involvement is needed to ensure an AI system can successfully play its envisioned role (we revisit and expand on these criteria later on).

Yet, while such criteria can help secure the place of various human contributors in the realm of discoverers proper, AlphaFold, too, may have a claim to making more than tool-like contributions. Its contributions are relevant, without them many protein structures could not easily have been discovered; they are not easily replaceable (unless we consider trivial variations of its code base as genuine rivals); and the processes enabling them are computationally expensive (Senior et al., 2019; Skolnick et al., 2021). So what role, exactly, does AlphaFold play? Let us consider an analogy to further explore the tensions AC experiences in categorising AlphaFold's contributions.

Consider Eric, who works as a research assistant on a project aiming to identify protein structures, much like AlphaFold. Assume that Eric knows nothing about biochemistry. Given some procedural instructions, he calculates the distances between

pairs of amino acids and works out the angles between their chemical bonds to model a protein structure. He then submits his results to the project's lead researcher, who assesses the significance of Eric's findings. Let us assume, favourably, that Eric has much more skill than other researchers at this particular task—they are not trained to perform tedious manual calculations, but Eric is, and he performs them diligently and swiftly due to his extensive experience. Without Eric, the team could not have identified protein structures, at least not close to the pace or level of accuracy matching Eric's. Is Eric a discoverer?

Even when Eric has no wider knowledge, no understanding of biochemistry and no interpretive abilities, it does not seem obvious that he is not discovering when he works out the three-dimensional structure of proteins. Conversely, however, it also does not seem right to name Eric a key discoverer. After all, it was not Eric who came up with the analytical and computational procedures he followed or who interpreted and applied his results. Turning to AI systems like AlphaFold once more, neither engineers or researchers developing and deploying AI systems, nor the interpreters of their results, seem to be the sole relevant target for ascribing discovery; but neither is AlphaFold or Eric.[7]

So if we were to treat both Eric and AlphaFold equally, as we think we should, which side should we lean towards? Should we consider both to make mechanical, tool-like contributions to discovery only, or grant that their contributions should be counted as discovery proper?

Perhaps the best response is neither of these options. It seems plausible to think that any scientific discovery can be decomposed into multiple ingredients, some of which are more important than others, and that we might draw a fuzzy boundary between more redundant, easily furnished, and non-autonomous contributions, and those significant, unique, and effortful enough to be considered genuine contributions to discovery. AC is conceptually unprepared, however, to recognise contributions that go beyond those of mere tools but stay shy of full-fledged discovery. On AC, being a discoverer is all-or-nothing. But both Eric and AlphaFold suggest that this is not always plausible. Before we turn to articulating a view that captures and systematises these intuitions, let us briefly consider a third option for applying AC that brings us closer to what we need, but not quite close enough.

3.3 Both humans and AI discover

A third variant of applying AC to AI discovery is to say that both the AI and the human researchers involved in a project are discoverers proper. This construal departs from narrower versions of AC that focus on identifying a *single* discoverer. However, a wider reading seems appropriate, as at least some authors think that there can be groups of discoverers, such as Stuart who suggests that "[t]he agent can be an individual or a community, whose mind can be extended or distributed" (Stuart, 2019, p. 52).

[7] An important difference between Eric and AlphaFold could be that Eric's contributions are especially meaningful for the attention, care, and sacrifice (e.g. of alternative uses of his time) they involve. We agree that these could constitute important disanalogies, so we ask our readers to imagine, for the sake of argument, an especially dispassionate but professionally diligent version of Eric. We thank an anonymous reviewer for suggesting ways to refine the analogy.

At first blush, this option makes some progress towards accommodating the concerns outlined earlier. But we think that it is not quite enough to be compelling. First, as elaborated previously, existing AI systems like AlphaFold still lack understanding and interpretive abilities thought to be essential for discovery on existing renditions of AC like Stuart's and Whewell's. Being more lenient when it comes to how many agents may count as discoverers leaves other tenets of AC untouched, such as that discoverers are those who impactfully use their abilities to understand, interpret, recognise significance, and so on. So unless emphasis on such abilities were suspended, it seems difficult to reconcile identifying both humans and AI systems as discoverers with central tenets of AC.

Second, to include AI systems among the set of discoverers, we still need clearer criteria for distinguishing the roles they play from those of mere tools (e.g. the hardware on which AlphaFold is running). Existing versions of AC do not offer resources for making such distinctions in a compelling manner, and would need to be substantially extended to provide a richer account of how discoverers could be identified among the messy background of tools and helping factors. We make some proposals for how such distinctions can be informed later when detailing our alternative to AC, but importantly, to work well for AC, such distinctions would also need to address the intrinsic tensions arising from AC's emphasis on a discoverers' ability to understand and interpret: what good is it if AlphaFold is robustly identified as more than a mere tool, but still lacks many of the abilities AC considers central for being a discoverer?

Third, while we think that, overall, recognising that AI systems can make more than tool-like contributions is important progress, simply counting them as part of a set of discoverers and putting everyone in this set on equal footing fails to recognise important nuances about the nature of different contributions made, how they matter individually, and how they interact to enable a discovery. In complex epistemic projects, like those involving AlphaFold, different agents and entities each make essential and non-redundant contributions to a discovery. But not all contributors seem to be discovering in the fuller sense expected by AC. Nor is it clear that all discover in the same sense, or, indeed, that any one individual or entity discovers by themselves. We articulate a view that captures these ideas more fully in the subsequent section. For now, let us briefly touch on how AC struggles not only with AI discovery, but also with recent and historical cases where only humans are involved.

3.4 AC struggles with human discoveries

As elaborated earlier, AC resists counting AI systems as discoverers for their inability to understand what they are discovering and to apply this understanding beyond the scope of an epistemic project. But paradigmatic cases of human-led discovery can exhibit similar features.

Consider Fleming again. Copeland (2018, p. 701) suggests that while Fleming played an important role in the discovery of penicillin, recognising its bactericidal effect on staphylococci, he did not, and perhaps could not have, proceeded from the realisation that penicillin was an antibiotic towards practically applying his findings. In the case of penicillin, important advances building on Fleming's initial discovery were

achieved by Ernst Chain and Howard Florey, alongside many others, who successfully identified non-trivial methods for purifying penicillin (Ligon, 2004). In light of this, at least part of the significance often ascribed to Fleming's findings is afforded by the practical import that was only later realised through efforts made by others, as is perhaps evidenced by Fleming sharing the 1945 Nobel Prize in medicine with Florey and Chain (Ligon, 2004, p. 52).[8]

Relatedly, there are reasons to think that it was not Fleming's original stroke of genius which discovered that penicillin prevented the growth of bacteria (Copeland, 2018, p. 706). Indeed, moulds had been used since antiquity to heal wounds (Wainwright, 1989) and British physiologist John Scott Burdon-Sanderson observed that bacteria did not grow in the presence of the fungus Penicillium almost 60 years before Fleming's finding (Keys, 1987). Moreover, as Copeland emphasises, following Chain (1971), Fleming's observations were not necessary for the discovery of penicillin, as there are reasons to believe that the discovery was 'in the air' and would have successfully been made by others regardless of Fleming's contributions (Copeland, 2018, p. 706, p. 716). Finally, Copeland argues that there was indeed a whole epistemic community and network of collaborators (albeit perhaps a poorly connected one) that Fleming was working within, and that played important roles in the discovery of penicillin, widely construed (2018, p. 701, p. 710, p. 717). These observations do not square up well with AC's tendency to allocate credit in an all-or-nothing fashion, and usually to single individuals.

Many other cases of discovery sans AI pose similar problems for AC. Some discoverers never understand the significance of what they find, such as in discoveries of elements like Lithium, whose electromagnetic properties were only explored a century later (Reddy et al., 2020). Many mathematical discoveries are also not made in a context of wider understanding or with an idea of how they may be used, but as isolated proofs. But even when researchers lack context or are unable to apply their findings, we still tend to think they are discoverers in an important sense, although perhaps not in the sense demanded by AC. Reinforcing similar points made by Copeland (2018), what such cases suggest is that AC is ill-equipped not just for understanding discovery involving AI, but discovery more generally. Moving beyond the critical contributions made by Copeland and ourselves, let us proceed to outline a view that helps overcome these limitations.

4 The collective-centred view

The collective-centred view (CC) maintains that focusing on a single agent or group of core collaborators is a misguided way of thinking about scientific discovery. Instead, CC insists the discovering entity is a *collective* in the first instance. The collective discovers *as a whole*, and in virtue of the diverse contributions made by its constituents. While this allows that single individuals and entities may discover in the way expected

[8] For lack of space, we will not address here the thorny issues of how prior and follow-on work should be credited (see Copeland, 2018 for an interesting discussion), but note that a refined version of our proposals should address these issues explicitly.

by AC, including by successfully deploying specific abilities to prompt useful observations, to interpret them, and to recognise their significance, it may also be the case that no single individual or entity discovers in this extensive way. Instead, only the collective as a whole discovers in virtue of how the contributions made by its constituents interact and complement each other.[9]

So in contrast to AC, CC maintains that for many (or most) scientific discoveries.

(1) *There is* **no clear discovering agent** *who conducted all, or at least the important part, of the discovery process*.
(2) *A* **collective of actors and entities all made important contributions to the discovery**.
(3) *Credit for discovery should be* **distributed between these agents** *depending on the nature and significance of their contribution*.

In emphasising these features of discovery, CC hence builds on a rich tradition of highlighting the interconnectedness of discovery processes and the often large, unwieldy networks involved in discoveries. For example, Kuhn (1970) has pointed out difficulties that prevent historians from attributing discoveries to any single person or point in time; Hollingsworth (2008) draws out a number of ways the discovery process is interconnected by presenting five different levels which contribute to the development of major discoveries; and Merton (1973) argues all discoveries are 'in principle' multiple discoveries. CC is also consistent with work on collective responsibility, such as Dang's, who rejects a unified concept of epistemic responsibility in science that "defines epistemic responsibility as an all-or-nothing concept" (Dang, 2018, p. 1). Instead, she proposes "epistemic responsibility should be distributed among members of a group when epistemic labour is distributed." (ibid.) Our account equally fits well with insights obtained from the study of distributed epistemic enterprises (e.g. Gargiulo et al., 2022; Huebner et al., 2017; Kukla, 2012). In cases like these, but also less heavily distributed epistemic projects, CC insists that discovery is best understood to happen at a collective level, with credit distributed between agents based on their contributions.

How do AC and CC compare and relate to each other more broadly? In terms of scope, CC extends significantly beyond AC. It handles cases that AC is conceptually unprepared to elucidate, including discovery involving AI, large-scale research projects, decentralised distributed epistemic projects, and historic discovery episodes that involved more widely distributed contributions on re-examination. So, at the level of what cases the two views are prepared to capture, we might say that CC is simply more general than AC.

But CC also makes several more substantive departures from, and improvements over, AC. First, as emphasised earlier, by putting the collective at the centre, CC is prepared to recognise a wider network of contributors *from the start*. CC urges us to begin from the assumption that discovery is likely to be distributed among different agents and entities, and helps us narrow the collective down, if and when appropriate. In virtue of this change in priors, CC departs substantially from AC, which is disposed

[9] We will not say much on interactions between different contributions to discovery here (e.g. how they amplify or inhibit each other, or complement and compound) as this is not essential to our aims and seems best explored in the context of concrete case studies.

to identify smaller groups, and fails to recognise the contributions made by agents that do not fit its restrictive conditions wholesale.

An immediate worry with this change of approach could be that CC places too much emphasis on the collective, at the expense of individual-level events. A second major departure from AC helps respond to this concern: CC is conceptually oriented towards understanding how the various contributions to a discovery made by different agents and entities can vary along several important dimensions and interact and compound towards enabling a discovery. As we elaborate shortly, CC indeed offers more scope to emphasise the role of individuals within a collective than AC, attributing credit to individuals based on the nature and significance of their contributions. CC is hence well-equipped to recognise that idiosyncrasies and specific backgrounds and skills of particular scientists help them make highly pertinent and non-redundant contributions (cf. Barwich, 2021) and that focusing on them can be essential for understanding discovery. CC can thus retain AC's emphasis on discoverers' abilities to understand and interpret, but, importantly, also insists that not all contributions must involve such abilities; it is enough that some do.

Finally, CC departs from AC through its ameliorative ambitions. It aims to help us more appropriately (re-)characterise historical discovery episodes, such as Crick and Watson's celebrated discovery of DNA. Here, more recent reconstructions have emphasised Rosalind Franklin's crucial contributions to the discovery, turning away from the prior focus on the heroic, lab-coated male discoverers as per traditional discovery templates (Klug, 1968; Maddox, 2003; Rapoport, 2002). In pursuing such ambitions, CC seeks to prepare the grounds for more just allocations of credit for contemporary and future discoveries.

While extending significantly beyond AC in terms of cases captured and goals pursued, the scope of CC must also be carefully delimited when it comes to questions answered. CC is not supposed to provide definitive guidance on settling practical issues, including (1) how to assign authorship, (2) how to distinguish authors from acknowledged contributors, (3) what practices are appropriate for recognising the role of prior work in enabling discovery, (4) how to recognise and celebrate discoveries with awards, (5) what resources beyond those afforded by the scientific publication system are needed to facilitate more nuanced allocations of credit for discovery, or (6) how CC bears on the incentive structures faced by researchers and their productivity. While these questions are highly relevant, particularly in view of CC's ameliorative ambitions, our aim here is to outline a general philosophical account that improves our understanding of how discovery is constituted. While CC is hoped to inform more specific, practical proposals addressing the issues highlighted here, it should not be expected to address them all by itself. Instead, CC is best understood as providing useful philosophical foundations that can support concrete attempts to understand discovery and to study, criticise, and improve existing practices of allocating credit for discovery across the sciences (see e.g. Dang, 2019; Kleinberg & Oren, 2022; Rescher, 2021; Rubin & Schneider, 2021).

With its central tenets and scope clarified, let us proceed to further detail CC, specifically focusing on the central questions of who makes up a discovering collective and what criteria may be used to distribute credit within the collective.

4.1 Making up the collective and distributing credit

Consider AlphaFold again. Who in the DeepMind team should be credited for the discoveries made with its help? Which contributions should be recognised as genuine bits of discovery and which only as tool-like contributions? Should we consider contributions by other actors within and beyond DeepMind but outside the AlphaFold projects, such as the funders enabling the research or the developers who wrote prior code which AlphaFold developers built on? And how should we credit even more remote contributions from computer science, mathematics, statistics, and other relevant fields which helped prepare the theoretical and technological grounds for AlphaFold's discoveries? Similar questions arise in other widely epistemically distributed collective discoveries, such as in biomedical research (Huebner et al., 2017; Kukla, 2012; Winsberg et al., 2014), or high energy physics projects at CERN, where papers often list thousands of team members, and projects are reliant on sophisticated technology, funders, and policymakers.

The challenge in fleshing out CC is essentially the same as that faced by AC when distinguishing between genuine bits of discovery and mere tool-like contributions: in order to avoid discovering collectives becoming impractically unwieldy and to meaningfully credit those who contributed significantly to a discovery, we need to draw a line somewhere.[10] A second challenge follows on the heels of casting a discovering collective in sharper outlines: how do we share out credit *within* the collective? Neither saying each contributor works alone nor that each discovery is made by everyone equally seems compelling. Even when distributing credit more widely, we still want to emphasise some contributions as more important than others, or else explain why this is not the case. It seems plausible, then, that contributions to discovery come in degrees and we should look to features of particular contributions to articulate how significant they are. To make progress on both challenges, let us build on our earlier suggestions and elaborate several features of contributions to discovery that CC can use to inform who may be included in the collective and how credit may be distributed.

4.1.1 Relevance

Relevance captures the idea that some contributions are more relevant to a discovery than others in a difference-making sense. Other things being equal, an agent who makes a more relevant contribution should more likely be counted as part of a discovering collective, and be awarded a larger share of the credit. Of course, assessments of relevance can be difficult to make, and much hinges on appropriate counterfactuals. For instance, funding can make a significant difference to enabling discovery, but are funders discoverers? Not plausibly. For one, we might think that their contributions, while highly relevant, lack other important features. Moreover, there might be other relevant counterfactuals than 'no funding, hence no discovery', such as a different funder playing the same role. This idea is captured by a second feature.

[10] While AI systems are the only technologies we explicitly consider, our account leaves space for other sophisticated technologies to make more than tool-like contributions. We do not explore this possibility here.

4.1.2 Redundancy

Redundancy concerns how essential a particular contributor's skills are or how replaceable a contribution is to a discovery. For example, the role of a desktop computer as a platform to write code on is highly redundant—there are many, easily available alternatives to meet the same functional role. On the other hand, a supercomputer along with software to enable distributed computation, may be much harder to replace and therefore stake a greater claim to being included in a collective. Once more, it also seems that the counterfactuals used to assess redundancy must be suitably constrained to allow for meaningful assessments: we should not assess redundancy with respect to faraway possible worlds, e.g. where supercomputers and highly skilled researchers can be found on every corner, but focus on whether contributions can be realistically substituted in close-by possible worlds.

4.1.3 Time/effort

A third important feature is how much time and effort is spent in furnishing a contribution. For example, if a graduate student spends 100 h pipetting, it can seem appropriate to recognise their contribution, even if they play a redundant and not highly skilled role. But like other features, significant time and effort spent alone may not be enough for inclusion in the collective, nor are time and effort always reliable metrics, such as when agents use their time inefficiently. At the same time, it is also important to recognise mere efforts towards discovery even when they do not directly pay off, such as when researchers divide labour between them but only one strikes gold.

4.1.4 Directness

Directness captures how directly a contribution is related to a discovery. For example, a highly skilled mechanic who is responsible for maintaining the cryogenic systems for cooling the LHC's electromagnets is not redundant—not many people would be able to perform her role, she may draw on knowledge and experience specific to proprietary components of the accelerator, and she might put large amounts of time and effort into her work. However, she might not be directly involved in any *specific* discovery, as her work is simply aimed at making the cryogenic system and the accelerator function more generally, regardless of what experiments are run on it. By contrast, engineers involved in designing a particle detector to function within a specific envelope required for particular collision experiments have a more substantial claim to making significant contributions, as figuring out the 'how' in practice can be as important as figuring out the 'what' in theory, and an effective contribution by an engineer is one that connects both in a specific, appropriate way.[11]

4.1.5 Originality

Discovery aims at, and partly consists in, identifying and understanding *novel* phenomena, concepts, and ideas, and finding better ways to meet our goals. Contributions

[11] We thank an anonymous reviewer for highlighting this point.

which promote these aims hence often involve some form of originality and, in virtue of this, seem to deserve more credit than those reusing existing approaches to familiar ends. Determining how original a contribution is, of course, comes with its own backpack of challenges and we do not suggest that doing so is ever easy: applying a new method to an old problem can be as original as the converse, and 'new' never guarantees 'better' or 'more impactful' (see Tin, 2003). So, cashing out originality requires significant nuance and much will hang on how we describe particular contributions. For example, reading results off a graph does not necessarily seem to merit a high originality score. However, just like many before and after Fleming have 'glanced at petri dishes', reading off results in a way that offers a unique interpretation to a problem can indeed make highly original contributions. Without attempting to provide a definitive guide to assessing originality, our suggestion here is that some measure of originality, sensitive to issues of context and description, should figure in a plausible appraisal of contributions to discovery.

4.1.6 Leadership and independence

Finally, we think it is important to consider how great a managerial role a contributor played and how independent their contribution is. For example, a lead researcher who envisions a research project, armed with suspicions of what promise it holds, and who designs processes, develops methodology, manages the project, and synthesises results from different branches, may deserve more credit than a researcher who is dependent on detailed, mechanical instructions to complete their contribution. Here, we might also think that *goal-directedness* plays an additional role—an AI system may make highly direct, but entirely un*directed* contributions for lack of understanding its own activities. The vision needed to conceive of a project and the managerial abilities involved in directing available means to aims worthy of pursuit deserve special attention.[12] Importantly, however, not only lead researchers can make contributions that exhibit these features; specialised junior researchers or engineers that are in touch with street-level challenges may often make contributions that score highly on leadership and independence too.

How do the features outlined here help determine who is part of a discovering collective and how credit should be distributed? First, all of them can be exhibited to different degrees, and being part of a discovering collective can be understood as a property that supervenes on these lower-level features, and possibly others not considered here. Second, none of the features are individually necessary or sufficient for being counted as part of a discovering collective or for being awarded a large credit share, regardless of the degree to which they are realised. A more plausible understanding is to take them as gradual conditions for *candidacy*: exhibiting none or only a few and/or only to small degrees may get you nowhere near being a part of a discovering collective. But there are various combinations that might strike us as

[12] To be sure, leadership and goal-directedness may play a relatively less important role in *serendipitous* discovery for lack of a clear goal to be pursued. However, as Copeland argues, even happenstance discoveries can be strategically promoted by cultivating skills to "[…] perceive value in unexpected observations and to utilise those observations strategically" (2018, p. 715), which might be understood as a form of goal-directedness.

warranting a sufficiently strong claim to be considered. This does not imply that there is ever a sharp boundary—many interesting and controversial cases will sit on a fuzzy border that separates those others we can place uncontroversially as within or without the collective. This is a feature, not a bug, since crediting contributions as genuine bits of discovery can be intrinsically controversial and substantive disagreement will often remain. Yet, while it seems elusive to formulate general strictures on which combinations of the features articulated here issue tickets to discovery shares, we think that emphasising their importance is a crucial first step for improving our understanding of discovery.

With the outlines of CC in place, let us revisit AlphaFold once more to see how CC can help assess its contributions.

4.2 AlphaFold is a candidate for the collective

As suggested earlier, AlphaFold makes highly relevant contributions to the discovery of protein structures: were it not for AlphaFold, many of these structures could not have been identified on a comparable schedule.

In terms of redundancy, AlphaFold does well, too. Not only is it highly specialised and expensively trained (Cheng et al., 2022), but there are also few institutions which can afford the amount of computing power needed to train AlphaFold, rendering the software and its concrete instantiation on specialised hardware highly non-redundant. Related considerations apply when it comes to time and effort spent: while vastly quicker than human computation, a large amount of time and computing power goes into the calculations AlphaFold makes.

Concerning directness, AlphaFold scores highly, too. Identifying features to predict how proteins fold marks a significant part of what constitutes the discovery, at least once complemented by a suitable interpretation. Unlike the LHC mechanic working on general tasks not directly related to any specific discovery, AlphaFold is directly involved in producing contributions that partly make up the discovery at issue.

Regarding originality, one might think that AlphaFold's contributions cannot be original as it merely follows its programming. This is true, but only in an unimportant sense: deep learning-based AI systems must exhibit considerable degrees of autonomy and flexibility to succeed at learning tasks, and learning outcomes can be highly novel and impossible to anticipate. AlphaFold's functioning centrally involves learning features that promote predictive success—and in doing so it can make original connections we have not made, and likely could not have easily made otherwise, which plausibly warrants a claim to originality.

Finally, concerning leadership and independence, it seems that AlphaFold may not score too highly. It does not set itself up to address real-world questions of interest to scientists, does not devise the research questions to be addressed or the methods for addressing them, and it generally lacks goal-directedness. While zooming in further could suggest that AlphaFold 'guides' discovery at least in the sense of generating a 'method' for predicting protein structures, e.g. by learning connection weights that help predict these structures, this is perhaps not enough for granting any significant claim to leadership. However, future AI systems may well be more independent in

choosing research trajectories and act in a more self-directed way, so it is far from clear that AI systems, in general, cannot also exhibit leadership and independence.

On several of the features that CC emphasises, AlphaFold is hence a strong candidate for being counted among the discovering collective, and for being awarded a share of credit. Interestingly, on a first assessment, AlphaFold could even be considered to deserve more credit for discovering three-dimensional protein structures than Fleming, our archetypal discoverer, does for his contribution to the discovery of penicillin. As noted by Copeland (2018, pp. 705–707), Fleming's observations were not necessary to the discovery of penicillin, since "[…] at least seven scientists prior to Fleming had noted the effectiveness of penicillin in inhibiting bacterial growth" (de Rond & Thietart, 2007, p. 548). So unlike AlphaFold, Fleming may not do too well concerning redundancy or originality.

How do AlphaFold's contributions compare to those of its human collaborators? This can be difficult to tell, partly since detailed information about how AlphaFold was developed is difficult to access, and partly because CC's resources for appraising contributions to discovery can only take us so far. Importantly, we should not think that there is always a clear, correct answer to who is part of a discovering collective or how, exactly, discovery shares should be distributed. Instead, (1) substantive disagreement is to be expected, (2) much will hinge on case-specific details that may remain unavailable, (3) those involved in discoveries are often in a better position than outsiders to determine who contributed what,[13] and (4) appraising the significance of specific contributions often requires understanding how they interact with and complement contributions by others (e.g. in the style of Mackie, 1965).

In light of these complications, our aims do not include issuing conclusive comparative judgments about discovery shares, but rather focus on offering a general framework that helps underwrite such judgments, given details of specific cases. That said, the features CC highlights for appraising contributions to discovery can help us think more carefully about what we value in a contribution; stepping back from unhelpful norms which may prevent some contributors from feeling entitled to credit where it is due while leading others to inflate the worth of their contributions.

Let us consider and reply to some additional concerns and challenges we have not touched on so far, and sketch out ways in which CC may be detailed further.

5 Concerns and challenges

5.1 Machines do not make contributions of their own

The first concern is straightforward: why should we think that machines make contributions of their own at all, rather than humans having a full claim to any epistemic successes on the grounds that they developed, initialised, refined, and deployed the machines, and interpreted their findings? Many of these human contributions are

[13] See Copeland (2018, p. 706) for examples of how Fleming and Chain offered more nuanced accounts than standard narratives of the discovery episodes they and many others were involved in.

causally prior to and necessary for machines doing their work, and so could be understood to simply encompass all of what machines subsequently help discover.

However, first, much like our parents and teachers do not have a full claim to our personal achievements on the grounds that their actions enabled them, it is at least not obvious that researchers and developers hold a *full* claim to the discoveries made with AI systems by virtue of guiding these systems to learn to predict successfully. Enabling is not the same as doing, at least with regard to those achievements that involve substantial autonomy on the part of the system in question. Second, in suggesting AI systems' contributions are entirely encompassed by their creators, humans are implicitly set up as the opposite: independent agents able to make distinct and autonomous contributions worthy of credit. But when pressed, this distinction seems difficult to maintain. Take AlphaFold: human researchers set up and carefully tuned AlphaFold to dispose it to learn how to predict protein structures well. But for AlphaFold to function properly, it must not only be designed in the right way and equipped with useful constraints on learning, but also afforded with sufficient autonomy to learn whatever there is to learn. More generally, the structure of deep neural networks is often unpredictable; they do not take in direct rule-based inputs; they often come up with surprising ways to relate inputs, intermediaries, and outputs; and they regularly produce results that their creators could not have anticipated (cf. Boge, 2022). The distinction between a *learning* algorithm, i.e. the human-specified algorithm by which a machine learns, and a *learned* algorithm, i.e. the specific, independently learnt way in which a machine relates inputs, intermediaries, and outputs, is useful here (Zednik & Boelsen, 2022, p. 233). While AlphaFold depends crucially on humans for its learning algorithm, and draws on further human supervision and guidance, its *learned* algorithm is not concretely determined by choices that researchers and developers make (ibid.).[14]

What is more, it seems plausible that if a human researcher, possibly guided by extensive instructions, were to come up with a similarly successful learned algorithm for predicting protein structures as AlphaFold, this would warrant a candidacy claim for inclusion in the collective, even if they did not further interpret or assess the significance of that algorithm, understand or grasp things about protein structures by using it, or exploit the algorithm's derivational resources. More generally, while many human researchers can be dependent on extensive training, detailed instructions, and continued oversight (consider Eric), it seems implausible to think that the contributions they make should be *entirely* attributed to those enabling and guiding their functioning as parts of a larger epistemic machinery. It is unclear, then, by what criteria a general

[14] To be sure, human researchers are of course heavily involved in promoting the system's ability to gravitate towards a useful learned algorithm, e.g. by making architectural choices or fine-tuning hyperparameters. Nevertheless, the way in which the learning algorithm interacts with the training data, and the relationships between amino acid sequences and three-dimensional protein structures encoded in these data, more concretely determines *what* structures AlphaFold ultimately predicts. Something similar may also seem true of instruments like particle detectors, or some models and simulations: their outputs are not trivially determined by human choices either. Intuitively, an important difference is that many AI systems can be described as 'learning' and 'making inferences', which involves degrees of autonomy not easily matched by scientific instruments that employ specific, human-devised mechanisms for detection and measurement, or models and simulations, whose derivational resources are specified by humans rather than formed independently. Despite these differences, it might be interesting to consider to what extent sophisticated scientific instruments may also exhibit the features that CC emphasises, and what this suggests about their roles in discovery.

distinction between machines and humans could be sustained regarding whether they can make contributions to discovery of their own, or whether credit is absorbed by those enabling and steering their roles in a discovery process.

To be sure, we do not seek to offer a definitive judgement here as to whether AlphaFold makes substantial contributions to discovery of its own—we expect and appreciate controversy even concerning our modest suggestion that AlphaFold is a *candidate* for the collective. Moreover, AlphaFold should not be taken as a definitive reference case of machines making contributions to discovery of their own, but only as a useful vignette to catalyse our thinking about AI discovery. But while we are sympathetic with those remaining sceptical of attributing significant discovery contributions to AlphaFold and similar systems, it seems plausible to think that future AI systems may exhibit higher degrees of autonomy, to the point where even those currently sceptical of ascribing discovery contributions to AI systems may be pressed to consider this option more fully.

5.2 Machines are inadequate targets for credit

A second concern follows closely behind the first. Even if we grant that machines can make discovery contributions of their own, we might nevertheless think that machines are not the kinds of things that credit for such contributions can be owed *to*.[15] Our practices of allocating credit and responsibility for discovery are partly supposed to express that we recognise and value someone's contribution; but AI systems may simply lack essential properties necessary for credit attributions to play this role.[16] Relatedly, many authors emphasise that credit allocations play important instrumental roles, for better or worse, e.g. by creating incentives for researchers to be productive or to make certain kinds of contributions rather than others (see Goldman & O'Connor, 2021; Wu et al., 2022; Zollman, 2018). On such a view, it seems that credit may be misallocated to machines, e.g. because they do not experience the same credit-related needs as humans, e.g. in applying for jobs and funding.

We have several replies here. First, allocating credit only to those able to appreciate it fails to pick out the right agents. Compare a key contributor, Carol, who routinely downplays her contributions and feels uncomfortable in the spotlight, and a less central contributor, Jess, who feels strongly (and perhaps appropriately) respected and valued to receive recognition. It seems odd to give more credit/reward to Jess on the grounds that they would appreciate it more. Relatedly, posthumous credit sits awkwardly in this picture of who credit may be owed to, pressing us to imagine counterfactuals of how much a recognition would be appreciated *if* an agent were able to appreciate it. Finally, there are numerous agents who we might think do not typically deserve discovery credit, e.g. student lab assistants, but whose contributions we nevertheless (should) respect and value, e.g. by paying them and/or thanking them. So, it seems unlikely that an account of who is suitably receptive to allocations of credit can plausibly sustain a broad distinction between humans and machines.

[15] We thank Jannik Zeiser for raising this concern.

[16] This is broadly related to the *responsibility gap* discussed in contexts of moral issues arising from AI applications (Matthias, 2004). We thank Jannik Zeiser for highlighting this similarity.

Second, ascribing credit to express that we value a contribution is just one among several functions that this practice plays. Others include signalling, e.g. 'attaching' contributions to agents to form a broader picture of their overall achievements, which, albeit very imperfectly, can help others assess how impactful someone's work is overall, how productive they are, and how innovative they might be in the future. It seems that similar functions could also be important in assessing and comparing the broader epistemic credentials of AI systems, even if allocating credit and responsibility to AI systems falls short of serving other functions.

Third, if the whole credit for a discovery that builds centrally on AI systems' contributions would be allocated only to humans, this would simply over-credit humans. Like allocating the credit for a well-written paper to a single author but not their ghostwriter seems problematic in many contexts, we must avoid over-crediting humans in a way that fails to track their true contribution to a discovery. So, while it may be right to say that we don't owe credit *to* AI systems, a sound practice of credit allocation may nevertheless involve making allocations to these systems in the ways suggested by CC, if and when indicated by the significance of their contributions.

Finally, throughout our discussion we have taken a view of credit as being due in proportion to the epistemic contributions an agent/entity has made. We agree that such a view may not always seem appropriate given the incentive-related roles that credit attributions play and that there are important questions about desert, need, and function arising when sharing out credit. At the same time, we also believe that we can usefully distinguish between an epistemic and practical layer of deciding where, and how much, credit is due. Perhaps a good amount of progress is already made by recognising that machines can make important epistemic contributions worthy of a form of credit that tracks these contributions, even while maintaining that credit-as-incentive should be distributed differently, taking into account additional reasons that favour humans.[17]

5.3 CC is intractable in practice

A third concern focuses on the practical difficulties with tracing out varied contributions and understanding how they interact to enable a discovery. Consider massively distributed research projects with hundreds of contributors playing different roles, like those at CERN, where it seems unlikely that we can readily assess the individual relevance of specific contributions, how original they are, how much effort they involved, how they productively complemented each other, and so on. Here, we might worry that, while CC's descriptive and ameliorative ambitions are laudable, it is just too complicated to track and process the features it emphasises for carving out the discovering collective and distributing credit shares.

We agree that applying CC to concrete discovery episodes can prompt important and sometimes insurmountable difficulties. But, as indicated earlier, CC is not supposed to precisely identify the weight of specific contributions, capture their interactions, and output a definitive score tracking an individual's or entity's discovery share. Rather, CC is supposed to promote a general attitude for thinking about discovery that improves

[17] We thank an anonymous reviewer for encouraging us to say more on this issue.

upon AC: it urges us to be open-minded about acknowledging a potentially wide range of contributions made by different actors and entities, and not to take for granted that discovery can be neatly ascribed to clearly identifiable discoverers, while relegating all else to the blurry and circumstantial background.

Moreover, the sorts of real-world complications that give rise to this worry are best understood as properties of the discoveries in question—CC does not induce these complexities, it just helps recognise them. Our dissatisfaction should hence lie, if at all, with the fact that complex phenomena like scientific discovery are difficult to understand in all their detail, rather than with a view that helps us recognise that discovery is often more complex than ordinarily considered.

More generally, then, we do not think that the practical difficulties arising in applying CC to concrete cases should make us reluctant to endorse it as a better alternative to AC. And while more work is needed to navigate important difficulties in helping CC elucidate concrete discovery episodes, it seems that even imperfectly precise recognition of contributions that would otherwise remain unacknowledged makes important progress over AC: incremental progress is progress after all.

5.4 CC makes credit distribution less just, not more

Finally, a fourth concern follows on the heels of the previous: applying CC in practice may end up making attributions of credit not more just, but less so.[18] For instance, we might think that the features CC offers for characterising and appraising contributions to discovery exhibit scope for abuse, such as when agents already in positions of privilege and power incorrectly describe their contributions as more significant than they really are, but these descriptions remain unchallenged for reasons related to the power differentials that enabled them. Or, less maliciously, we might simply think that positions of power often come readily equipped with in-built potential for making larger, and real, differences to discovery episodes, in non-redundant ways, and involving plenty of leadership. But some of these potentials are perhaps not credit-worthy, e.g. being highly relevant to a discovery because one has the power to shut down a project.

We are highly sympathetic to this concern, and think that a refined version of CC should be complemented by safeguards to avoid over-allocating credit to agents who are more influential mainly for the powers they can exercise, while downplaying the contributions of other, less privileged individuals. That said, CC as we have sketched it here is not a recipe that *guarantees* good results (e.g. just distributions of credit), but only a tool that can play a supporting role in such recipes. Like with any tool, it must be used in the right way to make an envisioned difference, there are real concerns about improper use, and we must ensure that other tools and resources we draw on complement it well. So, while we think that CC does not address this challenge out-of-the-box, we hope that detailed case-study-based work (following calls by feminist philosophers, see e.g. Richardson, 2010) can help further explore how CC can be complemented with useful safeguards that help equalise power differentials in negotiating

[18] We thank an anonymous reviewer for highlighting this concern.

membership in discovering collectives. Such a project is as important for CC as it would be for AC, which struggles with similar concerns.

With these challenges articulated and some responses sketched, let us conclude and take stock of what CC contributes to our philosophical understanding of scientific discovery.

6 Conclusions

Agent-centred views (AC) of scientific discovery assume that there is a neatly identifiable agent (or group of agents) usefully understood as the 'discoverer'. But despite what narratives of celebrated historical discoveries suggest, discovery processes are often more widely distributed and are not adequately characterised by picking out central actors responsible for a discovery. Focusing on scientific discovery involving Artificial Intelligence (AI) systems, and building on existing criticisms, we have argued that AC is an inadequate view not just of scientific discovery involving AI, but also of discovery involving only humans.

Moving beyond these criticisms, we have outlined a collective-centred alternative, CC, and argued that it is significantly better equipped to understand the complicated nature of modern discoveries, with and without AI, as well as giving us deeper insight into historical ones. Centrally, CC locates discovery in a collective, which can be constituted by humans and machines. Different agents and entities can each make important contributions to discovery, and CC offers conceptual resources to explore and appraise the significance of their contributions depending on various finer-grained properties.

Focusing on the role that AI can play in discovery, CC can reconcile existing enthusiasm about AI's promises with important hesitations about its limited abilities: AI systems might not discover in the extensive sense demanded by AC, but can nevertheless make significant contributions which must be recognised to appropriately account for the roles they can play in discovery.

Turning to discovery more generally, with and without AI, CC's richer conceptual resources can help us move considerably beyond AC's limited perspective: CC increases our ability to highlight why specific contributions stand out as centrally important for discovery, if and when they do, but also provides resources to acknowledge contributions made by agents and entities that AC would fail to recognise. In doing so, CC can support ameliorative projects aiming to promote more just distributions of credit for historical, contemporary and future discoveries.

Of course, CC also faces important challenges and limitations. For instance, it cannot precisely determine allocations of credit among a collective by itself, without rich, case-specific information. It also seems unclear how, exactly, CC can inform existing practices of distributing credit, assigning authorship, making acknowledgements, and so on. We believe that spelling out such details is important, but also that doing so is a larger project beyond the limits of this paper. While we think that our sketch of CC already makes important conceptual progress in refining our understanding of discovery, we hope that, much in the spirit of the view we defend, our proposals will be complemented by subsequent, collective efforts involving not only philosophers but

also sociologists of science and computer scientists to spell out CC's details further and examine its applicability and usefulness.

Acknowledgements We would like to thank the audience and organisers of the PT-AI 2021 conference for feedback on an early version of the paper and Jannik Zeiser for his insightful comments pushing us to consider further critiques of our view. We are also very grateful to two anonymous reviewers for their time and expertise offering comments which helped us greatly in improving the manuscript.

Author contributions Elinor Clark conceived of the main ideas for this paper and wrote an initial draft, which was enhanced and refined through discussions with Donal Khosrowi and submitted to Synthese. Upon receiving promising reviews and calls for further revisions and refinements, Donal Khosrowi was brought on board as co-author and both authors have since substantially reworked and refined the paper's ideas in close collaboration.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by Deutscher Akademischer Austauschdienst (Grant No. 91814091).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest, including affiliation with or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ardanuy, M., McDonough K., Krause, A., Wilson, D. C. S., Hosseini, K., & van Strien, D. (2019). Resolving places, past and present: Toponym resolution in historical British newspapers using multiple resources. In *Proceedings of the 13th workshop on geographic information retrieval (GIR '19)*. Article 3, 1–6. Association for Computing Machinery. https://doi.org/10.1145/3371140.3371143

Banerjee, A., & Monteleoni, C. (2014). *Climate change: Challenges for machine learning*. NIPS tutorial. Retrieved June 21, 2021, from https://www.microsoft.com/en-us/research/video/tutorial-climate-change-challenges-for-machine-learning/

Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., & Allen, C. (2021). Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews, 175*, 113–806. https://doi.org/10.1016/j.addr.2021.05.016

Barwich, A. (2021). Fishing for genes: How the largest gene family in the mammalian genome was found (and why idiosyncrasy in exploration matters). *Perspectives on Science, 29*(4), 359–387. https://doi.org/10.1162/posc_a_00375

Best, M., & Neuhauser, D. (2004). Semmelweis and the birth of infection control. *BMJ Quality & Safety, 13*, 233–234. https://doi.org/10.1136/qhc.13.3.233

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds & Machines, 32*, 43–75. https://doi.org/10.1007/s11023-021-09569-4

Brannigan, A. (1981). *The social basis of scientific discoveries*. Cambridge University Press.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*, 5339–5372. https://doi.org/10.1007/s11229-018-01949-1

Callaway, E. (2022). What's next for AlphaFold and the AI protein-folding revolution. *Nature, 604*, 234–238. https://doi.org/10.1038/d41586-022-00997-5

Chain, E. (1971). Thirty years of penicillin therapy. *Proceedings of the Royal Society of London: Series B, Biological Sciences, 179*(1057), 293–319.

Cheng, S., Wu, R., Yu, Z., Li, B., Zhang, X., Peng, J., & You, Y. (2022). FastFold: Reducing AlphaFold training time from 11 days to 67 hours. *arXiv*. Retrieved March 30, 2022, from https://arxiv.org/abs/2203.00854, https://doi.org/10.48550/arXiv.2203.00854

Copeland, S. (2018). 'Fleming Leapt Upon the Unusual like a Weasel on a Vole': Challenging the paradigms of discovery in science. *Perspectives on Science, 26*(6), 694–721. https://doi.org/10.1162/posc_a_00294

Dang, H. (2018). Epistemic responsibility in science. Social Epistemology Networking Event handout. Retrieved April 2, 2022, from https://www.haixindang.com/uploads/5/9/8/4/59847021/dang_new_handout.pdf

Dang, H. (2019). *Epistemology of scientific collaborations*. Doctoral dissertation, University of Pittsburgh.

de Rond, M., & Thietart, R. (2007). Choice, chance, and inevitability in strategy. *Strategic Management Journal, 28*, 535–551.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics, 37*, 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Fleming, A. (1929). On the antibacterial action of cultures of a Penicillium with special reference to their use in the isolation of B. influenzae. *British Journal of Experimental Pathology, 10*, 226–236.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gargiulo, F., Castaldo, M., Venturini, T., & Frasca, P. (2022). Distribution of labor, productivity and innovation in collaborative science. *Applied Network Science, 7*, 19. https://doi.org/10.1007/s41109-022-00456-0

Goldman, A., & O'Connor, C. (2021). Social epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 ed.). https://plato.stanford.edu/archives/win2021/entries/epistemology-social/

Hagendorff, T., & Wezel, K. (2020). 15 challenges for AI: Or what AI (currently) can't do. *AI & Society, 35*, 355–365. https://doi.org/10.1007/s00146-019-00886-y

Halina, M. (2021). Insightful artificial intelligence. *Mind & Language, 36*, 315–329. https://doi.org/10.1111/mila.12321

Hanson, N. R. (1960). Is there a logic of scientific discovery? *Australasian Journal of Philosophy, 38*, 91–106.

Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry, 1*(3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007

Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. In C. Hannaway (Ed.), *Biomedicine in the twentieth century: Practices, policies, and politics, volume 72 of Biomedical and Health Research* (pp. 317–353).

Huebner, B., Kukla, R., & Winsberg, E. (2017). Making an author in radically collaborative research. In T. Boyer-Kassem, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge: New essays.* Oxford University Press. https://doi.org/10.1093/oso/9780190680534.001.0001

Hull, D. L. (1988). *Science as practice: An evolutionary account of the social and conceptual development of science*. University of Chicago Press.

Jumper, J., Evans, E., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Keys, T. F. (1987). Beta-lactam antibiotics for clinical use. *Therapeutic Drug Monitoring, 9*(1), 126.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125*(5), 524–543. https://doi.org/10.1037/0033-2909.125.5.524

Kleinberg, J., & Oren, S. (2022). Mechanisms for (mis)allocating scientific credit. *Algorithmica, 84*, 344–378. https://doi.org/10.1007/s00453-021-00902-y
Klug, A. (1968). Rosalind Franklin and the discovery of the structure of DNA. *Nature, 219*, 808–810. https://doi.org/10.1038/219808a0
Kuhn, T. S. (1970 [1962]). *The structure of scientific revolutions* (2nd ed.). The University of Chicago Press.
Kukla, R. (2012). "Author TBD": Radical collaboration in contemporary biomedical research. *Philosophy of Science, 79*(5), 845–858. https://doi.org/10.1086/668042
Langley, P. (1977). BACON: A production system that discovers empirical laws. In *IJCAI'77: Proceedings of the 5th international joint conference on artificial intelligence* (Vol. 1, p. 344).
Ligon, B. L. (2004). Sir Howard Walter Florey—The force behind the development of penicillin. *Seminars in Pediatric Infectious Diseases, 15*(2), 109–14. https://doi.org/10.1053/j.spid.2004.04.001
Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.
Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*(4), 245–264.
MacLeod, B. P., Parlane, F. G. L., Morrissey, T. D., Häse, F., Roch, L. M., Dettelbach, K. E., Moreira, R., Yunker, L. P. E., Rooney, M. B., Deeth. J. R., Lai, V., Ng, G. J., Situ, H., Zhang, R. H., Elliott, M. S., Haley, T. H., Dvorak, D. J., Aspuru-Guzik, A., Hein, J. E., & Berlinguette, C. P. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*. https://doi.org/10.1126/sciadv.aaz8867
Maddox, B. (2003). The double helix and the 'wronged heroine.' *Nature, 421*, 407–408. https://doi.org/10.1038/nature01399
Magnani, L. (2000). *Abduction, reason, and science: Processes of discovery and explanation*. Kluwer.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1
Merton, R. K. (1973). *The sociology of science, theoretical and empirical investigations* (p. 356). University of Chicago Press.
Mills, C. (2007). White ignorance. In S. Sullivan & N. Tuana (Eds.), *Race and epistemologies of ignorance* (pp. 11–38). State University of New York Press.
Morgenstern, L. (2001). Mid-sized axiomatizations of commonsense problems: A case study in egg cracking. *Studia Logica, 67*, 333–384. https://doi.org/10.1023/A:1010512415344
Mullard, A. (2021). What does AlphaFold mean for drug discovery? *Nature Reviews Drug Discovery, 20*, 725–727. https://doi.org/10.1038/d41573-021-00161-0
Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science.* University of Minnesota Press.
Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Network, 113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012
Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM '18: Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 3–3). https://doi.org/10.1145/3159652.3176182
Qureshi, A. P., Stain, S. C., & Solomon, N. L. (2021). Diversity in scientific discovery. *The American Surgeon, 87*(11), 1732–1738. https://doi.org/10.1177/00031348211023411
Rafati, J., & Noelle, D. C. (2019). Efficient exploration through intrinsic motivation learning for unsupervised subgoal discovery in model-free hierarchical reinforcement learning. *arXiv*. https://arxiv.org/abs/1911.10164
Rapoport, S. (2002). Rosalind Franklin: Unsung hero of the DNA revolution. *The History Teacher, 36*(1), 116–127. https://doi.org/10.2307/1512499
Reddy, M. V., Mauger, A., Julien, C. M., Paolella, A., & Zaghib, K. (2020). Brief history of early lithium-battery development. *Materials (Basel), 13*(8), 1884. https://doi.org/10.3390/ma13081884
Reichenbach, H. (1938). *Experience and prediction. An analysis of the foundations and the structure of knowledge*. The University of Chicago Press.
Rescher, N. (2021). Allocating scientific credit. In *Ethics matters*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-52036-6_14
Richardson, S. S. (2010). Feminist philosophy of science: History, contributions, and challenges. *Synthese, 177*(3), 337–362.
Rubin, H., & Schneider, M. D. (2021). Priority and privilege in scientific discovery. *Studies in History and Philosophy of Science, 89*, 202–211.

Sap, M., Shwartz, V., Bosselut, A., Yejin, C., & Roth, D. (2020). Commonsense reasoning for natural language processing. *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020

Schaffer, S. (1986). Scientific discoveries and the end of natural philosophy. *Social Studies of Science, 16*(3), 387–420. https://doi.org/10.1177/030631286016003001

Schiebinger, L. (1987). The history and philosophy of women in science. *Signs, 12*(2), 305–332. https://doi.org/10.1086/494323

Schiller, F. C. S. (1917). Scientific discovery and logical proof. In C. J. Singer (Ed.), *Studies in the history and method of science* (Vol. 1, pp. 235–289). Clarendon.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., & Hassabis D. (2019). Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). *Proteins, 87*, 1141–1148. https://doi.org/10.1002/prot.25834

Shorter, E. (1984). Ignaz Semmelweis: The etiology, concept, and prophylaxis of childbed fever. *Medical History, 28*(3), 334.

Skolnick, J., Gao, M., Zhou, H., & Singh, S. (2021). AlphaFold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *Journal of Chemical Information and Modelling, 61*(10), 4827–4832. https://doi.org/10.1021/acs.jcim.1c01114

Stuart, M. (2019). The role of imagination in social scientific discovery: Why machine discoverers will need imagination algorithms. In M. Addis, F. Gobet, & P. Sozou (Eds.), *Scientific discovery in the social sciences.* Springer.

Tin, T. B. (2003). Creativity, diversity and originality of ideas in divergent group discussion tasks: The role of repetition and addition in discovering 'new significant', or 'original' ideas and knowledge. *Language and Education, 17*(4), 241–265. https://doi.org/10.1080/09500780308666851

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence, 4*, 189–191. https://doi.org/10.1038/s42256-022-00465-9

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., & Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research, 50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Wainwright, M. (1989). Moulds in folk medicine. *Folklore, 100*(2), 162–166. https://doi.org/10.1080/0015587X.1989.9715763

Whewell, W. (1849) Of induction, with especial reference to Mr. J. Stuart Mill's system of logic. John W. Parker.

Whewell, W. (1996 [1840]). *The philosophy of the inductive sciences* (Vol. II). Routledge/Thoemmes.

Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radically collaborative research. *Studies in the History and Philosophy of Science, 46*, 16–23. https://doi.org/10.1016/j.shpsa.2013.11.007

Wu, J., O'Connor, C., & Smaldino, P. E. (2022). The cultural evolution of science. Preprint retrieved July 20, 2022, from https://doi.org/10.31222/osf.io/2ekcr

Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., & Oda, K. (2016). A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science, 351*(6278), 1196–1199. https://doi.org/10.1126/science.aad6359

Zednik, C., & Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds & Machines, 32*, 219–239. https://doi.org/10.1007/s11023-021-09583-6

Zollman, K. J. S. (2018). The credit economy and the economic rationality of science. *Journal of Philosophy, 115*(1), 5–33. https://doi.org/10.5840/jphil201811511

Zytkow, J. M. (Ed.). (1996). *Machine discovery*. Kluwer Academic Publishers.