ABSTRACT

Title of Dissertation:	CONSTRUCTING OUR MORAL WORLD: AGENCY, TELEOLOGY, AND KORSGAARD
	Andrew Thomas Fyfe Doctor of Philosophy, 2023
Dissertation directed by:	Professor, Samuel J. Kerstein, Philosophy

Kantian ethicists maintain that morality applies to all agents irrespective of an agent's particular circumstances, interests, or concerns. That is, morality applies to an agent categorically rather than hypothetically. Kantian ethics attempts to prove this categoricity by deriving morality from the constitutive conditions of action. If such an argument could be made to work, then morality would follow from the constitutive preconditions or "logic" of agency and thereby apply categorically to all agents regardless of unique eccentricities concerning an agent's particular circumstances or interests. As a result, an argument for Kantian ethics typically adheres to the following formula: (1) providing a theory of agency that (2) entails that all agents are committed to a Kantian ethical outlook.

My focus in this dissertation is one of these arguments on offer for Kantian ethics. Specifically, the argument of Christine Korsgaard. I cannot fully defend her

argument here in its entirety, but with this dissertation I hope to provide the background work developing the necessary theory of agency in order for Korsgaard's argument for Kantian ethics to succeed. Specifically, I aim to put forward, develop, and defend the sort of non-standard, teleological theory of agency upon which Korsgaard's argument for Kantian ethics crucially depends. Moreover, with this dissertation I aim to attack the more widely accepted Davidsonian, causalist theory of agency which Korsgaard's Aristotelian-Wittegenstienian-Anscombian teleological theory of agency opposes and I argue we should adopt instead.

CONSTRUCTING OUR MORAL WORLD:
AGENCY, TELEOLOGY, AND KORSGAARD

by

Andrew Thomas Fyfe

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Samuel Kerstein (*Chair*)
Assistant Professor Harjit Bhogal
Professor Patricia Greenspan
Professor Dan Moller
Associate Professor Erich Sommerfeldt (*Dean's Representative*)
National Institutes of Health Bioethics Faculty David Wasserman

Acknowledgements

I must begin with my advisor Samuel Kerstein. Sam has graciously allowed me the freedom to pursue a dissertation with a wide-ranging scope that draws upon work from philosophy of biology, action theory, as well as debates regarding libertarian freedom – all in pursuit of accomplishing nothing less than a proof of Kantian morality. When I proposed this dissertation my committee complained that I was being too ambitious. They were correct. But under Sam's direction my committee allowed me the freedom to explore and pursue my ambitious aims despite their initial concerns because of their trust in me and I deeply appreciate that confidence and enabling willingness.

I wouldn't have had the opportunity to be here writing these acknowledgments without the generous support of Patricia Greenspan, Dan Moller, and Jerrold Levinson. Pat in particular has been my friend, steadfast advocate, and mentor throughout. I am also in debt to the rest of the wonderful faculty here at the University of Maryland for all their support, advice, and knowledge. In particular: Rachel Singpurwalla, Eric Pacuit, Peter Carruthers, Paul Petroski, Allen Stairs, Christopher Morris, Harjit Bhogal, as well as David Wasserman next door at the bioethics department housed within the National Institutes of Health. I am also thankful to the faculty at the University of Washington from whom I learned so much during my time there: Arthur Fine, Laurence BonJour, Marc Cohen, Robert C. Coburn, Stephen Gardiner, William J. Talbott and — in particular — Angela M. Smith. I must also thank a number of faculty at the University of Miami for my time there — Amie Thomasson, Peter J. Lewis, Risto Hilpinen, and Susan Haack — as

well as the faculty who welcomed and worked with me during my time as a visiting fellow in Harvard's philosophy department — Selim Berker, Samantha Matherne, and Christine Korsgaard (about whom I will have more to say about below).

To my philosophy mentors from undergraduate Charles Pailthorp, Joseph Tougas, and David Marr along with my physics faculty Thomas Grissom and E.J. Zita. I am eternally thankful to E.J. Zita for the kindness she provided me when after our first day of class she drove me to buy a blanket following a very cold first night back in college after the war when I had nothing but the clothes on my back.

To my fellow graduate students Christian Tarsney, Aiden Woodcock, Christopher Masciari, and Julius Schönherr whose work ethic as fellow graduate students has been an inspiration, whose feedback has always been penetrative, and whose friendship I will eternally fail to understand what I've done in life to deserve. Julius Schönherr has the peculiar honor of offering me my favorite criticism during this dissertation process, complaining that my dissertation was "building a coat hanger out of diamonds" in that each stage he found insightful but to the Kantian conclusions he wholeheartedly rejected.

I also owe a deep debt of gratitude to all the rest of my friends and fellow graduate students at the University of Maryland: Michael McCourt, Andrew Knoll, Kyley Ewing, Heather Adair, Jimmy Licon, Logan Fletcher, Ryan Ogilvie, Lane DesAutels, Quinn Harr, Javiera Maximiliana Perez-Gomez, Shen Pan, and Samuel Warren. As well as my graduate student co-travelers during my time at the University of Washington: Patrick Taylor Smith, Benjamin Visscher Hole IV, Elizabeth Scarbrough, Ali Hasan, David Alexander, Monica Aufrecht, Mitchell Kaufman,

Jeramy Gee, Aaron Hebble, Jon Rosenberg, Bennett Barr, Cheryl E. Fitzgerald, and the late Tyrel Mears whose further contributions to philosophy we have all sadly missed out on. I am also grateful to the Harvard philosophy department and all of the graduate students there who warmly welcomed me during my fellowship in Cambridge. But in particular: Florence Bacus, Britta Clark, and Zachary Gabor.

To my supportive friends outside of academia including Louise Gilman, Morgan and Emily Stinson, Daniel Cox, David Kahler, Darren Quintinson, Elle Mullins, Jane Sheehan, Vikki Connell, Michelle Bell, Gayle Wright, Marc Phillips, Jedidiah Sooter, Darryl Barkhouse, Kay Byfield, Josh Kinney, Heather Lindner, Andy Clark, Victoria De Capua, Andie Coller, Alex Mazzoni, Kasey Erb, Katie Keelan, Sean Mullen, Renata Rollins, Ina Newton, Kyle Pomykala, Amy Johnson, and Lindsey Fyfe. As well as my brother Taylor Fyfe and the endless reserve of support I've received from my parents Jeanne Dusel-Fyfe and David Fyfe.

To my supportive friends and colleagues inside of academia including Eric Brown, Paul Schofield, Liz Goodnick, Alex Hughes, Ian Halloran, Kris Wright, Marta Alvira-Hammond, Deborah Jean Weeks, Rachel Buchanan, Theresa Lopez, and Dila Gümüş. As well as Cindy Phillips who picked me up from the airport upon my arrival at the University of Maryland and continued to be a friend and colleague throughout our time here; Aida Roigé Mas who I picked up from the airport upon her arrival at the University of Maryland. Aida and I immediately began debating philosophy and neither of us ever had the good sense to cease. It's fitting that Aida and I would end up successfully defending our respective dissertations just days apart.

Damla Özakay for her boundless joy and laughter which reinvigorated my love of life when — I will admit — this process had beaten me down. To Kelsey Gipe of whom I couldn't even begin to know what to say about and so thereof I can only be silent. Bana önce "dost" kelimesini, sonra da dostum olarak anlamını öğreten Zeynep Maden'e.

Lastly, to Chris Korsgaard. I began my philosophical career working in the areas of epistemology, American pragmatism, and philosophy of science. I was youthfully skeptical of the work being done in ethics. Then one summer I read *The Sources of Normativity*. Immediately upon finishing I had the thought, "Wow. This is correct. I do this now." Since that day my philosophical work has been devoted to making sense of Kantian ethics in a broadly Korsgaardean-spirit. I am deeply grateful to her not just for opening my eyes to the Kantian nature of our shared moral world but also for the many hours she graciously spent discussing philosophy with me in her office during my Harvard fellowship.

Table of Contents

List of Abbreviations

References to and citations of Christine Korsgaard's major works are given parenthetically in the text, using the abbreviations below, and for monographs citing the relevant section number whereas page numbers are used for essay collections. For example: (SN §1.1.1) for monographs and (CKE 179-83) for essay collections. For Korsgaard's uncollected essays and other works, I cite using the standard conventions I employ for all other citations throughout this dissertation.

Monographs

FC Fellow Creatures: Our Obligations to the Other Animals (2018)

SC Self-Constitution: Agency, Identity, and Integrity (2008)

SN The Sources of Normativity (1996)

SPR The Standpoint of Practical Reason (1981) (dissertation)

Collections

CA The Constitution of Agency (2008)

CKE Creating the Kingdom of Ends (1996)

Introduction

Kantian ethicists maintain that morality applies to all agents irrespective of an agent's particular circumstances, interests, or concerns — morality applies to an agent categorically rather than hypothetically. Kantian ethics attempts to prove this categoricity by deriving morality from the logic or constitutive conditions of action. If such an argument could be made to work, then a commitment to morality would follow from the very constitutive preconditions of agency itself and thereby apply categorically to all agents regardless of any unique eccentricities. An argument for Kantian ethics typically adheres to the following procedural formula: first, provide a theory of agency and then, second, prove how according to such a theory of agency all agents are inescapably committed to the Kantian ethical outlook whenever they act.

My focus in this dissertation is one of these arguments for Kantian ethics. Specifically, in my dissertation I am concerned with the work of Christine Korsgaard. While I cannot fully develop and defend her argument in its entirety, with this dissertation I hope to provide the background work necessary to establish the underlying theory of agency upon which Korsgaard's argument for Kantian ethics depends. Which is to say, I aim to accomplish the first of the two-step procedure for constructing a full defense of Kantian ethics by putting forth and defending the sort of theory of agency from which Korsgaard's argument for Kantian ethics can be launched.

To begin with, what do I have in mind by "Kantian ethics"? And what is Korsgaard's argument for Kantian ethics? I will take each of these questions in turn. To first illustrate what Kantian ethics comes to, consider a case of minor immorality:

> **Mike from IT's Haircut.** Mike is a fairly well-off IT professional. One of his friends tells him about a local barber who is on the brink of bankruptcy. In order to boost sales, this barber is slashing prices to win over new clients. Frugal by nature and in need of a haircut, Mike decides to go to this barber. On his way into the shop, Mike notices a large amount of firefighter paraphernalia around the interior of the shop and infers that he might get a further discounted haircut if he pretends to be a fireman. What's the worst that could happen if Mike's lie gets found out — disapproving faces? Mike is shameless in this regard and he'd still get his haircut. In the end, Mike decides to lie and is able to secure himself a haircut on the house.

All plausible moral theories would agree that Mike acts immorally. Nevertheless each will give a different account as to why and what is wrong with Mike's lie. What a Utilitarian would have to say about Mike is that his action brings about the lesser good rather than the greater good. The barber needs money more than Mike does. In the barber's hands, the money would have (presumably) gone further to adding to the total happiness in existence than the happiness created by Mike lying and keeping the money (because the barber is in a more desperate situation).

Kantians have a different take on Mike. The problem with Mike's lie does not reduce to the balance of goodness and badness added or subtracted from the universe, the problem is that in lying to his barber, Mike disregards his barber's will. What a Kantian (like myself) would have to say about Mike, is that Mike's action treats the barber as nothing more than a mere object in the world to be manipulated for Mike's own purposes rather than as an equally valuable free agent whose choices possess the same value as Mike's own. This is not a rigorous statement of Kantian ethics, but this example provides us with a workable idea of its spirit which will allow us to proceed. But now the question arises, "Why might we think this Kantian view of morality is true?" Christine Korsgaard offers the following argument for the Kantian ethical outlook:

> **(1) Guise of the Good.** Agents must take the freely chosen ends for which they act to be good (i.e. choice-worthy);
>
> **(2) Constructivism**. When agents take the freely chosen ends for which they act to be good, this cannot be due to the ends they pursue possessing an intrinsic goodness which the agent recognizes their ends to have but can only be because the agent takes their choice to bestow goodness upon their ends;
>
> *Therefore from premise 1-2:*
>
> **(3) Half-way to Kantianism.** When an agent acts, an agent is committed to taking the choices of *at least* themselves to bestow goodness (i.e. choice-worthiness) upon their chosen ends;

(4) Universalization. If I take my choice to bestow goodness upon my freely chosen ends, then I am committed to taking the choices of others as similarly bestowing goodness upon their freely chosen ends;

Therefore from premise 3-4:

(5) Kantianism. When an agent acts, an agent is committed to taking the choices of themselves *and all other agents* to bestow goodness (i.e. choice-worthiness) upon their chosen ends.

If the above argument is correct, then the choices and pursuits set by the barber's will should be respected and valued by Mike in some equivalent sense to his own. That is, the will of the barber and Mike must be treated consistently in Mike's decision making. When Mike lies to the barber to get his haircut at a discount, he is failing to treat the will and ends of his barber as equivalent to his own will and ends given how Mike's lie bypasses the will of the barber in order to accomplish only his own chosen aim of obtaining a free haircut.

Yet it's unclear whether Kantian ethics is correct or whether Korsgaard's argument in favor of Kantian ethics is the right way of proving such a moral outlook. I take the most problematic premise in Korsgaard's argument in favor of Kantian ethics to be "Universalization". As others have pointed out — including Korsgaard herself — if we stopped at the half-way point in the Kantian argument we'd be left with something similar to Jean-Paul Sartre's existentialism wherein agents must value themselves and their own free choices but are not thereby committed to also respecting the equal value of other free wills. It is the universalization premise that

gets us from Existentialism to Kantianism and I suspect that the entire fate of Kantian Ethics rests upon the success or failure of this specific universalization premise — and it is very unclear whether the universalization premise succeeds. After all, why does the practical necessity of an agent like Mike treating himself as bestowing value-choiceworthiness on his own chosen ends entail a commitment on Mike's part to treating all other agents as also bestowing value-choiceworthiness on those other agents' chosen ends (e.g. the barber)? To answer this question we must look to the nature of agency generally and human agency specifically.

With this dissertation I intend to defend the sort of view of human agency which underlies Korsgaard's argument that human agents must value agency universally in order to even act as individuals who value their own humanity and are thereby capable of action at all – what I've called the universalization premise. To this end, I've adopted the 'covering concept' model for my dissertation. This dissertation format consists of three oversized papers on related topics. Specifically, all three of the following papers are offered as an incremental fleshing out of the theory of agency upon which I take Korsgaard's argument in favor of Kantian morality to depend.

In *Life, Teleology, and Agency* I defend an Aristotelian-Wittgensteinian-Anscombian teleological theory of agency in contrast to the Davidsonian causalist theory of agency. The traditional Davidsonian account struggles with a number of seemingly fatal objections. I begin this first paper with what has come to be known as the "Problem of Deviant Causal Chains" for causalist theories of agency. In this paper, I rely on the Problem of Deviant Causal Chains as a jumping off point for

introducing and defending an alternative teleological Korsgaard-inspired theory of agency. This account of agency has the seeming drawback of relying crucially on teleological notions of which one may (appropriately) be skeptical. Therefore I devote most of this paper to making sense of a Korsgaardean teleological theory of agency using the two main candidate theories for a naturalistic understanding of telos offered to us by contemporary philosophers of biology: (1) the etiological or "selected-effects" account of functions most associated with the work of Karen Neander (1991) and Ruth Milikan (1984); (2) the ahistorical systems or "causal role" account of functions most associated with the work of Larry Wright (1973) and Robert Cummins (1983). The aim of this paper is to show how with a modified version of the ahistorical systems or "causal role" account of teleology we can and should come to view all living organisms as agents capable of intentional, teleologically justified actions due to the special self-reflexive nature of a living organism's teleological organization.

In *Human Life, Teleology, and Libertarian Free Agency*, I turn from agency broadly to human agency narrowly. In the traditional Davidsonian causalist theory of action, to be an action a movement must be caused by a mental state that also rationalizes it. However, what the traditional Davidsonian account strangely leaves out from its story of an action's cause is the most important element of action: *the agent herself*. In the traditional causalist theory, it seems to be the states within the agent rather than the agent herself which are the cause of a movement. This is what has come to be known as the "Problem of the Disappearing Agent". As an alternative, it is sometimes argued — following Harry Frankfurt (1971) — that to be an action a

movement must not only be caused by a rationalizing mental state but, additionally, that the agent must "identify with" or "endorse" her being moved by the rationalizing mental state. This is one way of re-introducing the agent into the story of an action's cause. However, such Frankfurt-inspired alternatives really just push the problem back a level since we are still left having to make sense of an agent performing the mental act of identifying or endorsing. Mustn't the agent also have to identify with her mental act of identification? Endorse her endorsement? An infinite regress looms. In this paper I defend an alternative account of agency which doesn't fall afoul of the Problem of the Disappearing Agent and — in particular — I address how human, free agency proves to be unique and in a sense "deeper" than the agency of plants and non-human animals.

In *Diachronic Agency as Group Agency and Korsgaard's "Practical Identities"* I provide an account of temporally-extended or diachronic agency. What does it mean for me to decide today about what to do tomorrow? Let's begin by exploring what it cannot mean. Deciding for tomorrow is not merely doing something today which controls my actions tomorrow. It isn't like there is a button in front of me – which if I push – ensures that tomorrow I will robotically be controlled to carry out a certain plan of action. When I decide to do something today for tomorrow, there is still the agency of myself tomorrow that must freely choose to carry out the prior decision. But this poses a problem for the standard model of future decision-making. It is held by Davidsonian causalists that decisions for the future involve the creation of the mental state of an "intention" which — if it survives until the appropriate time — will then control my actions at that future time. So a decision today for what to do

tomorrow involves the creation of an intention which *controls* my actions tomorrow and seemingly "crowds out" my decision-making and control tomorrow. But such an account shares a troubling similarity with the scenario where I push a button today which simply robotically controls my movements tomorrow. The problem is that such an account cuts out the decision-maker tomorrow in a way that our actual decisions for the future do not. Maybe the problems with these theories of diachronic agency can be fixed, maybe not. There are also other proposals for making sense of future decision-making which I consider and reject as inadequate for various reasons in this paper, but I won't give them all here. The conclusion I ultimately work towards is Korsgaard's suggestion that we try and make sense of diachronic agency in terms of joint agency through the employment of Korsgaard's notion of "practical identities". This proves to mirror a view being advanced by game theorists like Natalie Gold regarding what's termed "team reasoning" — although with some important differences which I will address. Specifically, one way of understanding what it means to decide for tomorrow is that it involves the myself-of-today and the myself-of-tomorrow deciding *jointly* about what temporally-extended joint course of action to carry out *together* as a *team*. The idea here is that our diachronic, extended agency – the capacity to decide for later times – depends upon our deliberating, deciding, and acting jointly along *with* our future selves. It's not that we literally talk and negotiate with our future-selves in order to together agree upon a joint plan of action – after all I can't speak with my future self – but instead that myself today and tomorrow carry out our deliberative decision-making in a way that involves us doing what we'd hypothetically be able to agree upon in negotiation between our past, current, and

future self. We might envision this being performed by our engaging in the same sort of joint reasoning that we often do with strangers when we carry out group actions like deciding on a time to meet for lunch through a shared, collaborative deliberative process. In fact, diachronic agency is really just a sort of group agency which arises from the different momentary time-slice deliberators occupying our body serially (à la Derek Parfit) deliberating and working together as a team to set joint ends and perform group actions.

Collectively, the three papers which make up this dissertation outline an account of agency which make it clear why we might think there is a plausible further argument to be made from agency to Korsgaard's "Universalization" premise. This ultimately accomplishes the first step, but not the second in the two-step formula for grounding Kantian ethics: first, provide a theory of agency and then, second, prove how according to such a theory of agency all agents are inescapably committed to the Kantian ethical outlook whenever they act. I leave the second step to later work. However, to see how the account of agency I develop in this dissertation serves the needs of Korsgaard's argument for Kantian ethics, let's return to Mike from IT once more before closing out. The mistake Mike makes is failing to take his barber's will as bestowing his chosen ends with value just like Mike's own. Mike pursues his own chosen ends in a way that disregards any "buy in" from the barber to the scheme that Mike has opted to pursue and involve the barber in. This violates the Kantian dictate that we treat the humanity in ourselves and others as the source of value rather than as just another object that possesses value insofar as it is useful to us. But why is Mike from IT committed to Kantian ethics? Why must Mike value and respect not only his

own choices but those of his barber as well in a way that rules out lying to his barber and thereby bypassing his barber's will? There are two routes of advance we might pursue building upon the account of agency I develop throughout this dissertation. While I leave pursuing these two further routes of advance to later work, I'll briefly preview both here.

The first line of argument goes like this: Korsgaard argues that when human agents take their own choices as bestowing value, what they are doing is valuing their own "humanity" as a generic trait which possesses the "normative standing" to bestow value upon that agent's freely chosen ends. However, no agent is able to take themselves to possess such normative standing on the basis of their generic humanity without also being committed to all other free agents also possessing the same power to bestow value upon their freely chosen ends. Valuing one's humanity is like possessing the right to vote due to one's place of birth. An agent cannot claim that one possesses the right to vote on the basis of one's place of birth without recognizing the equal right to vote of anyone else sharing this same place of birth. The upshot is that an agent like Mike cannot consistently value only his own humanity when he acts without additionally being committed to also valuing humanity generally. This includes the humanity of himself *and all others* whenever he acts and takes his own humanity to possess such normative standing. That is, an agent cannot value only one's own humanity and take only oneself to possess the power to bestow ends with choiceworthiness while denying that any other similarly free agent's choices possess this same power or "normative standing".

However, it might be wondered why humanity is a trait that can only be valued in this "generic" manner — rather than only something that provides me with normative standing but not others. The analogy or metaphor with voting is all well and good, but is humanity actually the sort of trait that can only be valued in oneself in this manner? Korsgaard gestures at a possible answer here which I build upon further in this dissertation:

> Couldn't he, that is, decide to respect only his own humanity? This is an ill-formed question. What is your own, in the individual sense of your own, is not your humanity but what you make of it, your [personal] identity, and the existence of that depends on your respect for humanity in general. (SC §9.7.4)

The first two papers making up this dissertation develop an account of humanity which sets up exactly such an argument to be made. At the close of the second paper, I will have established how our humanity - our capacity for self-determination - does turn out to be a species trait unlike other features that make up our unique personal identities. Although humans share the same human form given by nature, the generic human form is precisely the trait of being the sort of life who must create their own specific form to become a unique individual agent with their own personal identity. Other forms of life - blades of grass, spiders, horses - instantiate a form of life that's exactly the same as any other instance of the same species. Human life is distinctive in how each individual selects for oneself a unique form of life to live and embody. As Korsgaard points out: "[A] non-human animal's life is mapped out for it by its instincts; and any two members of a given species basically live the same sort of life (unless the differences are biologically fixed, as by age and gender, or by kinds as among bees)." (CA 142). Which is to say, our humanity isn't something we possess individually, but rather a generic species trait.

Human animals are specifically those sorts of living organisms who are self-determining in a particular self-aware way that involves one choosing one's own way of life. As a generic trait I possess as the type of being I am, it isn't the sort of property I possess individually in a way that makes it possible to value the capacity for self-determination in oneself without being rationally committed to also valuing this capacity generally and so in others.

The upshot is, that our humanity is a special sort of species general trait which perhaps it is impossible to value only in one's own instance unlike the other sorts of features of one's life and identity. For other things one might be able to value them just in one's own case, but one's humanity is shared generally in a way that can't be so individually valued due to its unique metaphysical place in our constitution. In this way, our relationship to our own humanity is distinctive from how we relate to other features making up our identity.

It isn't until the third paper in this dissertation that I turn to setting up an account of diachronic agency which can be used to underlie the second approach to arguing for Korsgaard's "Universalization" premise. This second approach echoes a line of reasoning mostly associated with Henry Sidgwick's *The Method of Ethics* (1874) and Thomas Nagel's *The Possibility of Altruism* (1978). As our point of departure for coming to a quick understanding of what this approach amounts to, let's begin with a fairly natural way to think of the problem of ethics: I have desires, which give me reason to do what satisfy them. You have desires, which give you reason to satisfy them. But my reasons don't seem to give you any reasons nor do your reasons seem to give me any reasons and yet ethics more-or-less amounts to escaping the

selfishness of only pursuing the satisfaction of my own desires and instead giving the desires of others some weight. Another way of putting this conception of the problem of ethics is this: How do we justify to agents trafficking in private reasons the rationality of treating reasons as public?

This way of conceiving of ethics can make it seem as if a prudential sort of egoism wins by default unless morality can come along and talk the rational agent into caring about the reasons of others. But, as pointed out by Sidgwick and Nagel, the idea of reasons as private doesn't naturally lead to a prudential egoism - but rather to present-aim theory. Why is this the case? Well, it is unclear why one's distant future self's reasons should have any special value for you now distinct from the reasons of any other person. This opens up the advocate of private reasons and prudential egoism to what I'll term an "unstable middle attack".

An "unstable middle attack" is one in which you pose a target theory as an unstable middle ground between two different extremes which it can't successfully resist at the same time. For example, political libertarianism is open to just this sort of unstable middle attack. Libertarianism is flanked by anarchism on the one side and a view that embraces an expansive welfare state on the other. If you are a believer in inviolable individual rights, then you seem to get anarchism; if you believe that you don't have any absolute right to your stuff and we can tax it away for the least well off, then you come to an expansive egalitarian welfare state. But then this leaves the libertarian occupying a weird view in the middle which allows for a state where you can order people around and take their stuff — *but only a little bit*. In this way libertarianism doesn't partake in either of the advantages of the views it is flanked by

and so can't be maintained. It can neither cite inviolable individual rights to fend off the egalitarian view without collapsing into anarchism nor can it abandon absolute individual rights in the face of the needs of others and egalitarian considerations without collapsing into an expansive egalitarian welfare state.

Egoism is also open to an "unstable middle attack". We might imagine egoism as flanked by present-aim theory on one side and then an altruistic theory which gives consideration to others on the other. If you ask yourself why you should pay attention to your future self? Why should you save for your retirement and not gamble it away for the thrill of the risk or the immediate payoff that you might be able to spend now rather than wait for investments to grow in value slowly over time? If the egoist attempts to avoid present-aim theory by appealing to the impartial significance of your future self, you begin drifting into considerations which could be extended into an argument for the significance of other people. If the egoist attempts to avoid admitting impartial reasons by appealing to the way in which your desires have a special pull or reason giving force for you that other people's desires don't possess, then we open ourselves up to the charge that the desires of our future self are just as distant and alienated from us as the desires of others. Therefore, consistency demands that we not just become prudential egoists but rather present-aim theorists.

This provides the egoist with a dilemma: either (1) be a present-aim theorist or (2) abandon egoism. The egoist line of reasoning collapses to mere present-aim theory and any attempt to justify a concern for your future self only ends up continuing farther to eventually include all other agents as well. To evict the egoist from her unconcern for others, then, all there is left to do is to rule out present-aim

theory (i.e. once present-aim theory is ruled out, all that is left is the line of reasoning which results in a rejection of egoism). I aim to develop an account of diachronic agency which can be used to launch an argument for this Sidgwickian-Nagelian approach to defending the universalization premise in the third paper of this dissertation.

By briefly overviewing these two routes of defense for the universalization premise I merely aim to gesture at how the account of agency developed in this dissertation can be (hopefully) made to work and serve as the jumping off point for a further future argument aimed at proving the crucial universalization premise in Korsgaard's argument for Kantian ethics. Korsgaard herself has at several points in her collective works attempted to make such an argument, however even Korsgaard's admits that a fully satisfactory case is yet to be made. My hope is that with a better understanding of the underlying Aristotelian-Wittgensteinian-Anscombian non-causalist teleological theory of agency upon which Korsgaard's argument depends, it will be possible to make progress toward fulfilling the promise of finding the elusive completed proof of a Kantian supreme principle of morality.

Paper 1: Life, Teleology, and Agency

§1. Introduction

In this paper, I defend an anti-causalist view of action according to which agents are self-reflexive teleological systems. Mere inanimate objects behave in ways that can only be explained causally; living organisms — qua self-reflexively goal-directed systems — also have a *telos* in their self-maintenance in virtue of which their movements inherit a rationalization or purposiveness and thereby come to constitute *actions* rather than just *mere behaviors*.

The account of agency I defend in this paper is Aristotelian and Wittgensteinian-Anscombian in spirit, but it is most directly an extension of a similarly inspired account of agency developed by Christine Korsgaard. My focus in this paper will be to, first, introduce the Aristotelian-Wittgensteinian-Anscombian teleological account of agency as it has been further developed by Korsgaard and then begin by showing how this view contrasts with the more widely accepted Davidsonian causalist view of agency. In particular, my concern will be to embed Korsgaard's account of agency in the contemporary philosophy of biology literature on naturalized teleology — something Korsgaard herself eschews. Teleological notions have — for good reason — fallen out of favor in both modern science and philosophy and so the way in which Korsgaard's views about agency depend upon teleology require an account of how these teleological notions can be naturalistically legitimated and understood within contemporary philosophy of biology.

In this paper my focus will not be the free agency of human beings specifically but rather the broader account of agency possessed by all "life" — i.e. self-reflexively goal-directed systems. Which is to say, the account of agency I will defend in this paper may range broader than some people will find natural. On the account of agency I defend, plants, mice, and human beings will all qualify as agents (while inanimate objects like rocks or artifacts like mice traps, cars, and carburetors will not). This is because on the broad notion of agency, all living things are capable of intentional movements of the right sort to constitute an "action" at least in the loosest sense of that notion. Intentionality — at least in the sense connected to judgements of success and failure (i.e. the very sense relevant to agency and action) — are not a matter of having a mental cause from which it inherits intentionality, but rather is a matter of a movement or effect having a purposive place in a teleologically organized system. And if the kind of intentionality relevant to action is a matter of a movement or effect having a purpose in a teleologically organized system rather than a mental cause, then much more besides things with minds can produce intentional movements and effects (at least in this sense of intentional). Korsgaard makes this point nicely in her own work:

> [I]t makes most sense to see the kind of intentionality that characterizes human action as being at the extreme end of a continuum. At the bottom of the continuum is simple intentional movement [...] we ascribe to plants [...]. At the next level are the movements of simple animals whose movements are guided by perception. [...] At the next level, perhaps, would be the movements of [intelligent] animals that have some idea of what they are doing or trying to do. And finally [...] there would be the movements of [reflective human agents.] (SC §5.4.7)

In order to make sense of the intentionality of actions we must move beyond traditional causalist accounts of agency and action. Such traditional causalist accounts

require actions to inherit their intentionality-purposefulness from the intentionality of their mental state-act cause. Human agents do enjoy a more narrow and unique form of agency that is free, self-determining, and self-conscious which some would want to reserve the honorific title of "action" and "agency" for labeling. I will however resist discussing that special case of human agency in detail until a full paper can be dedicated to it, but however you feel comfortable using the labels "agent", "action", and "agency" that is going to be — on my account — just a special case of the very general account of teleological movement which I will be defending in this paper and which also applies in its broader usage to both plants and non-human animals — that is, all life.

§2. The Problem of Deviant Causal Chains

When a rock comes tumbling down from atop a hill and we ask why the answer will come in the form of a causal explanation. However, in asking why an agent raised her hand we might be after an answer of a very different sort. Specifically, we might be asking for an intentional or teleological explanation; we are asking for the reason for why she raised their hand rather than the cause.[1] For example, intentional explanations like 'she raised her hand in order to ask a question' or 'to wave down an approaching friend' seem like satisfactory teleological explanations for someone asking why an agent raised their hand in most contexts. The

[1] Throughout I use the terms "teleological explanation", "intentional explanation", and "rational explanation" interchangeably. In the literature in philosophy of action these phrases are also typically used interchangeably. Additionally, I follow Davidson here in using the term "rationalizes" not in its now common derogatory sense where someone confabulates reasons to make an irrational action seem as-if it had been performed for good reason. Instead, I follow Davidson's use of the term "rationalizes" here to just mean "provide rational justification" or "make a movement rationally explicable by citing justifying reasons" or "rationally intelligible".

actions of agents are distinctive from the movements of inanimate non-agents like rocks in that actions are subject to intentional explanations instead of — or at least in addition to — mere causal explanations. The actions of agents 'are intentional' or 'have reasons' or 'have a purpose' whereas the movements of rocks never do.

And now a question: What can or should we say about the relationship between these two different types of explanation? One is appropriate for accounting for the behaviors of rocks rolling downhill — i.e. causes, causal explanations — and the other is associated with explaining the actions of agents — i.e. reasons, intentional explanations. According to anti-causalist theories of action, explanations which give reasons for actions and explanations which give the causes for mere behaviors of inanimate objects are wholly independent and *sui generis*. Which is to say explanations for actions are not causal, but irreducibly teleological; to explain an action, one cites the agent's reasons for acting — perhaps to be made sense of in terms of the agent's goal or purpose in performing the action — rather than giving the action's causes. Moreover, on the anti-causalist view an agent's reasons for acting are not what caused them to act. According to this view reasons are not causes nor is intentional explanation a species of causal explanation but rather an irreducible teleological form of explanation. The upshot seems to be that a complete explanation of the world cannot be given in purely efficient causal terms because the world includes agents and actions and these additionally require a different teleological form of explanation in terms of reasons.

Prior to Donald Davidson's essay "Actions, Reasons, and Causes" (1963) the consensus among philosophers of action was anti-causalist. But after Davidson's

notably influential 1963 paper this consensus within philosophy of action shifted. A new widespread and longstanding consensus formed favoring instead the causalist view that reasons are causes.[2] Behind the rapid acceptance of Davidson's new alternative causalist approach to action was many philosopher's concern that anti-causalism was at odds with a naturalist outlook on the world. A commitment to naturalism is — at least in part — usually taken to entail a rejection of teleological explanations (or at least a rejection of teleological explanations that cannot be "naturalized" by being reduced to a form of causal explanation after all). Davidson, for example, takes the only form of explanation to be causal and so insofar as intentional and teleological explanations are not causal they are epiphenomenal and not explanations at all. Davidson argued that "justifying reasons that do not cause us to act are rationalizations with no causal influence and, consequently, no explanatory value." (D'Oro and Sandis 2013:25).

Nevertheless, causalists like Davidson do not deny that actions are subject to teleological reason-focused explanations in a distinctive way that the behavior of rocks are not. And so Davidson proposed the following now paradigmatic form which *mutatis mutandis* all later causalist theories of action would take.[3] The standard view

[2] One reason for this shift was Davidson's argument that an anti-causalist view of reasons seemed unable to distinguish between genuine rational explanations for why an agent performed an action versus the rationalizations which merely rendered "intelligible why the agent *might* have done something without explaining why they actually did it" (D'Oro and Sandis 2013:25). Davidson argued that "justifying reasons that do not cause us to act are rationalizations with no causal influence and, consequently, no explanatory value." (D'Oro and Sandis 2013:25). The standard view became that actions could be casually explained by a mental state or event (such as a belief-desire pair or intention) which also justifies or "rationalizes" the action.

[3] Although it is worth noting: "Other causalists since Davidson are not wedded to the details of his account, but the basic idea remains the same: a common sense psychological explanation of human action cites (implicitly or otherwise) an appropriate mental state of the agent as the cause of the behavior (see, for example, Mele 1992 and 2003 and Bishop 1989; see also many of the essays collected in Aguilar and Buckareff 2010, as well as essays in D'Oro 2013)." (Sehon 2012:352).

within philosophy became that actions are to be casually explained by a mental state or event (such as a belief-desire pair or intention) which also justifies or "rationalizes" the action (i.e. teleologically explains the action). The causal theory of action claims that an action is a bodily movement with a special sort of cause. (D'Oro and Sandis 2013:18). What makes a movement or effect an action is being caused by the right sort of mental state (e.g. a belief-desire pair or intention), specifically one that can teleologically and intentionally explain the action. This is why actions turn out to be subject to intentional explanations in a way that the behavior of rocks are not. And *a fortiori* on this view the reasons cited by our rational explanations of actions are themselves efficient causes. This is why — in slogan form — Davidson's approach to philosophy of action became summarized as the view that "reasons are causes".

But not long after Davidson's 1963 essay that prompted the causalist wave, it was recognized that causalist theories faced a problem with deviant causal chains (Chisholm 1966; Taylor 1966; Davidson 1980). Donald Davidson presented a classic example of this sort in his paper "Freedom to Act" (1973):

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. (Davidson 1973:79)

In the above illustration, a belief-desire pair of the climber's both (a) causes the climber to let go and also (b) "rationalizes" his behavior (by "rationalizes" in this literature it is merely meant that it is explained by justifying reasons; in other contexts "rationalizes" often has a derogatory implication which the term doesn't carry in this

context), and yet still it does not seem that his "letting go" in this case constitutes an intentional action on his part. This is the problem of deviant causal chains.[4]

Many causalist attempts have been made to respond to this problem (Mele 1992; Peacocke 1979; Bishop 1989; Stoecker 2003). However, an increasing number of philosophers have come to share the sentiment expressed by George Wilson that once we truly comprehend the full force of the problem of deviant causal chains, "the evidence points to more than infelicity or incompleteness in the various causalist proposals – it points [...] to a global breakdown [of] the whole project" (Wilson 1989:258). The continued failure of the causalist theory of action to offer a satisfactory answer to the problem of deviant causal chains suggests that the theory is a failing "research programme" we ought abandon.[5]

However, there may still remain a satisfactory answer to the problem of deviant causal chains out there to be discovered or vindicated. However, I argue that no such yet-to-be-discovered answers are out there to be discovered and that, in fact, the problem of deviant causal chains is unresolvable and fatal. To see why, I will

[4] To restate the problem: The only criteria for a movement to be an intentional action according to the causalist theory is that it be caused by a mental state which provides a rational explanation for its performance. In the climber case the hand release is caused by a mental state which does provide a rational explanation for why the movement might have been performed, nonetheless the hand release doesn't seem intuitively to have been an intentional action or done by choice. So causalism gets the answer wrong in this case and similar cases. The solution is for causalist theories to add some additional requirement that the mental state cause which rationalizes the movement to cause the movement "in the right way". However, after decades of attempts to flesh out a satisfactory addendum to causalist theories of action which include this "caused in the right way" clause, there has been frustratingly little progress. So while the problem of deviant causal chains posed by the climber case may on its face seem like an easily addressable problem for causalist theories via simply adding a clause to the theory that requires a 'caused in the right way' addendum, in practice this has proven to be a perennial and stubborn problem that causalists have been unable to make progress in answering.

[5] Here I have in mind the work of Imre Lakatos (1970) who argues that broad paradigm theories are never refuted by "crucial experiments" but rather just prove themselves over a period as unable to handle objections and contrary evidence until we come to realize they are slowly degenerating "research programmes" and eventually abandon them.

begin this paper by offering Christine Korsgaard's argument that the failure of causalist theories to handle the problem of deviant causal chains to date is no accident but rather a symptom of a fundamental misunderstanding such causalist theories have about the nature of the intentionality of actions. As such, the failure of causalist theories of action to solve the problem of deviant chains over the last decades is no accident but an unsolvable problem resulting from a fundamental misunderstanding of intentionality and action embedded at the heart of causalist views of agency.

§3. Intentionality as Teleology

To be an action, a movement must be viewed as the agent trying to do something — an action is done for a certain purpose. This is why actions can succeed or fail, whereas causes never succeed or fail. Korsgaard expresses this point by saying that an agent's movements are always subject to a standard of efficiency. When we act, there is something the agent can be judged as trying to do and *a fortiori* judged to succeed or fail at doing. This efficiency requirement on actions, Korsgaard argues, is in fact what philosophers mean when they claim that actions must possess intentionality:

> [A]n action has intentional content. We say that spider is ''crawling to the center of her web to eat the moth that is trapped there'' or that the cockroach is ''running under the toaster to avoid being swatted.'' Those phrases specify purposes [....] The important thing about action is only that it is done, as we say, ''on purpose.'' To assign intentional content to a movement in this sense is to make it subject to a normative standard of efficacy, to a standard of success and failure. And being subject to such a standard, as I mentioned before, is essential to the idea of action. Suppose all we say about the cockroach is that ''he is running under the toaster.'' This is still different from saying of a rock that it is rolling down the hill. If the rock runs into an obstacle and stops rolling before it gets to the bottom of the hill, it has not failed. But if the cockroach does not make it under the toaster, he has failed. [...] [An agent performs an action when] his movements are subject to a standard of efficacy,

> a standard of success and failure. He is subject to a standard of success and failure, because there is something he is trying to do[.] (SC §5.4.4)[6]

On the typical account of the intentionality of actions, movements possess intentionality because they are caused by intentional mental states or acts (e.g. caused by belief-desire pairs, intentions, decisions). On the traditional Davidsonian-causalist view, actions are thought to inherit their intentionality from the intentional content of their mental state or mental act causes. Korsgaard rejects this. On her view, for a movement to possess intentionality is for it to relate in the right sort of way to a system possessing a structurally organized purpose. I'll have more to say about this idea in a moment.

The problem with the traditional causalist account where movements inherit their intentionality from their mental state causes is that such an account proves both too broad and too narrow. It is too broad in that it treats certain movements or effects as actions which shouldn't be classed as action. For example, the issue with over broadness is captured by "the problem of deviant causal chains" (albeit in a form slightly different from the one I raised before). I may desire letting my crush know that I am smitten by her, but believe it would be inappropriate for me to do so for whatever reason and so decide against it or at least still be thinking over whether to do so. Nonetheless, my desire might cause me to blush or stutter or be awkward

[6] What about actions performed purposelessly? Korsgaard answers: "Speaking roughly, and putting the point in a way best suited to human agents, an agent is efficacious when she succeeds in bringing about whatever state of affairs she intended to bring about through her action. A reminder is important here. I do not mean in saying this to imply an act is never done for its own sake; that is, I am not suggesting that an agent must always be trying to accomplish some purpose beyond the act itself. You might, for example, dance for the sheer joy of dancing. But even someone who dances for the sheer joy of dancing is subject to a standard of efficacy, because he can fail. He can, for instance, fall flat on his face. And if his steps are not in any way guided by this standard of efficacy—if he makes no effort not to fall flat on his face—then he is not dancing, but merely flailing about." (SC §5.1.3).

around her and in this way come to let my crush I am smitten by her. My blushing, my stuttering, and my awkwardness is the result of my desire to let my crush know I am smitten by her just as the action of telling her would be, and yet none of these constitute actions. Perhaps they possess intentionality in the sense that they express my emotions or mental attitude, but these effects are not intentional in the way actions are intentional. I cannot fail to blush or, for that matter, succeed at it. Whereas on the traditional causalist account of agency my blushing would constitute a fully intentional, purposeful action because it was caused by an intentional mental state with the aim of letting my crush know I am smitten by her just like how raising my hand to ask a question inherits its purposefulness from being caused by the intention to ask a question. While I don't intend to let my crush know I am smitten by her at this moment by blushing, because my blushing is caused by my intention to tell my crush I am smitten by her this makes my blushing intentional according to the Davidsonian causalist account of agency and intentionality. The causalist account gets things wrong in this case by counting my blushing as intentional when it is not.[7]

[7] Korsgaard discusses this problem herself — but not under the more common heading of "the problem of deviant causal chains". Rather, Korsgaard discusses this broadness problem as a matter of how emotional expressions can possess intentionality *of a sort* without possessing the kind of intentionality that actions possess — i.e. the sort of intentionality which makes a standard of efficiency applicable. Korsgaard uses as her example weeping. You might weep when you are hurt or at the news that your rival wasn't killed in an accident as it had first seemed and in both of these cases your weeping may possess intentionality in a sense. We can say you are weeping because you are hurting not just as a causal explanation, but in a way that makes your weeping intelligible; and we can morally reprimand you for weeping upon learning your rival survived because that reveals something about your attitudes. But in neither case are these actions even if they inherit "something like intentionality" from the mental states that cause them. The problem is that the sort of intentionality here isn't the sense related to success and failure: "When you say that someone is weeping, you do not invoke the possibility of failure that you do when you say that the cockroach is running under the toaster. What would count as failing? Of course you can try to weep, and fail, but if you try to weep and succeed, weeping to that extent is an action." (SC §5.5.3).

Besides being too broad, the traditional causalist account of intentionality is also too narrow. Intentionality in the sense that makes a standard of success and failure intelligible isn't unique to humans or even conscious agency. Arguably, not only humans are capable of intentional movements. Even plants might produce movements or effects which can succeed or fail. Korsgaard makes a similar point — although I will ultimately not be accepting the idea that artifacts are capable of intentional movements all by themselves:

> [T]o assign intentional content to an object's movement we do not need to suppose that there is some thought process that accompanies the movement. [...] Intentional descriptions apply even to the movements of artifacts. A clock, for example, is an object functionally constructed so as to tell the time. And that is why when we say things like ''This clock chimes out the hour'' or ''The alarm clock will wake me up at eight,'' we imply the existence of criteria of success and failure. If the clock chimes eleven times when both hands point to twelve, or if the alarm does not go off when eight comes round, then the clock has failed. It is because a clock is organized so as to tell the time that we can assign intentional content to its movements. (SC §5.4.5)

Intentionality — at least in the sense connected to judgements of success and failure, i.e., the very sense relevant to agency and action, is not a matter of having a mental cause from which it inherits intentionality, but rather is a matter of a movement or effect having a purposive place in a teleologically organized system.[8] And if the kind of intentionality relevant to action is a matter of a movement or effect having a purpose in a teleologically organized system rather than a mental cause, then much more besides things with minds can produce intentional movements and effects (at least in this sense of intentional).

[8] We find a similar point being made by Phillipa Foot concerning how the act of eating isn't just the mechanical movements by how those movements fit into the teleological structure of a way of life: "Eating, for instance, is essentially, conceptually, related to nourishment, and could not be conclusively identified by a story about the taking in, crushing, transforming, and spewing out of substances since, for all that, its purpose might be not the maintenance of tissue[.]" (Foot 2003:28).

The upshot is that movements which are subject to rationalizations are not just the movements of human agents, but the movements of anything teleologically structured. This leads to the Korsgaardian claim that the intentionality of movements — in the sense relevant to agency that produces a standard of success and failure — is a product of a movement belonging to a teleologically structured system. Korsgaard argues: "We regard [your lungs, your kidneys, your heart, your car engine, and your alarm clock] as capable of success and failure insofar as we regard them as functional objects, objects with some sort of job to do." (Korsgaard 2014:196). But this brings us headlong into the philosophically fraught issue of teleological functions. What does it mean to say your lungs, your kidneys, your car engine, and your alarm clock are teleologically functional objects with some sort of job to do? Or for that matter, what would it mean to say a thing is "teleologically organized"?

In order to answer these questions, in the next two sections I will explore the two theories of function and naturalized teleology that are prominent within contemporary philosophy — historical and ahistorical accounts:

> (1) [T]here are historical theories, which define functions in terms of the origins of the entity being functionally characterized. On a leading version of this account (defended by Ruth Millikan (1989) and Karen Neander (1995), among many others), the function of a heart is to pump because hearts' pumping blood in the past explains (in good Darwinian fashion) why hearts are around today. (Ridge et al. 2011:369-70)
>
> (2) [T]here are ahistorical theories, which define functions in terms of a system's capacities. Robert Cummins (1975) defended a classic theory of this sort, according to which functions are understood in terms of the contributions that parts of a containing system make to the capacities of that larger system. A function of the heart is to pump blood, because by pumping blood the heart enables the containing system (the organism) to survive and flourish. (Ridge et al. 2011:369)

After explaining the strengths and weaknesses of both (families of) theories of function, I will explain why either option would fall short from working for the purposes of the theory of agency I've laid out here. A theory of agency where the intentionality of a movement is not inherited from its mental state cause but instead from the place of that movement in the teleological life of an agent. Thus, in the rest of the paper I will build upon one of the two accounts — specifically, Robert Cummin's ahistorical, causal role theory of function — in order to provide an account which does the work to flesh out the theory of action I've introduced once it is combined with an Aristotelian account of life. This account, I will argue, has the virtue of combining the strengths of both families of views. But first I will begin by explaining in the next two sections why the traditional two proposals of function and naturalized teleology fail to work for the purposes of the theory of agency I've laid out here.

§4. The Selected-Effects Theory of Functions

Teleological terms such as "function" and "design" appear frequently in the biological sciences. Examples of the sorts of teleological claims made throughout biology include:

- A function of stotting by antelopes is to communicate to predators that they have been detected.
- Eagles' wings are designed for soaring.
- The kidney has the function of eliminating waste products from the bloodstream.
- The jackrabbit's big ears function to keep its body cool.

But post-Darwin, the persistence of teleological terms in biology is puzzling. Prima facie Darwin's theory of evolution seems to entail the illegitimacy of employing teleological notions in the biological sciences. Without an intending Divine artificer designing creations with a certain intentional purpose, the idea of viewing naturally evolved organs, animals, and traits as possessing teleology seems like a mistake. However, there is a faction within biology and philosophy of biology who argue that our tendency towards making teleological claims about nature isn't merely the continuation of a pre-Darwinian bad habit, but rather that function in a naturalized sense has value as a scientific term even within a modern, Dawinian biology. The motivation for the development and defense of a naturalistic account of teleological notions in biology comes from their persistence and apparent explanatory power within science. Proponents of the legitimacy of teleological notions in biology correctly point out that even among working biologists it is common for functional explanations to be set forth and accepted as satisfactory explanations. Suppose a biologist were to find the stotting behavior of antelopes strange and perplexing. Why do antelope do this? The explanation they seem to be after is information about the

stotting behavior's function; the functional explanation that an antelope's stotting communicates to predators that they have been detected. Similarly, is it not a scientific question why jackrabbits have such big ears? To be given the teleological explanation that they possess the function to keep the jackrabbit's body cool would seem to constitute a satisfactory answer to this scientific question.

But there is a concern to be raised here. The explanations as to why an antelope's stotting behavior and a jackrabbit's big ears are both vulnerable to the charge they are only metaphorical at best and misguidedly anthropomorphic at worst. The only explanation for the existence of these traits is as the products of the blind, purposeless process of evolution and natural selection. To treat as an explanation for these traits a claim that they possess a certain function does no real explanatory work and needlessly risks falling back into old teleological patterns of thought that invite one to confuse natural selection with intentional design. In order to clarify the problem I am pressing here, suppose we simply accepted by way of stipulation that it were true in some sense that the big ears of jackrabbits possess the function of regulating their body temperature. Even if this were true, how is this supposed to be explanatory? After all, a jackrabbit did not intentionally choose big ears in order to keep its body cool. While the function of big ears in regulating a jackrabbit's body temperature may be a plausible explanation for why a jackrabbit would opt for big over small ears (given the choice), big ears are not a product of a jackrabbit's own choosing. So, even if it were true that big ears had this function it would do nothing to explain the presence of the trait in jackrabbits. Short of positing a divine creator or engaging in a misguided anthropomorphization of the process of natural selection, we

cannot construe the big ears of jackrabbits as being the product of authorial discretion. Consequently, even if we granted that it were true that the trait of big ears in jackrabbits possessed the function of keeping a jackrabbit's body cool, it wouldn't have any explanatory power in regard to the question of why jackrabbits have such big ears.

However, defenders of teleology in biology contend that concerns like those above are misplaced. They argue that while it is the case that a jackrabbit's big ears are not the product of intentional selection, the trait is nonetheless the product of Darwinian selective processes. Suppose that in my studies of jackrabbits I can find no other significant role for their big ears in the lives or well being of jackrabbits except for how their large surface area serves to dissipate heat. Furthermore, I note that given the environment, way of life, and anatomical structure of jackrabbits they are uniquely prone to overheating. From these observations, I may be inclined to make the functional claim that the big ears of jackrabbits possess as their function cooling the jackrabbit's body temperature. The explanatory value of this functional explanation for why jackrabbit's have big ears will be fully naturalized and Darwinian. The trait of big ears in jackrabbits currently exists because the past ancestors of the trait (i.e. ears of progressively bigger size) were successful in enhancing fitness because they performed the function of cooling a jackrabbit's body, leading to the reproduction of the genotype for big ears. And consequently, if someone were to inquire after me why jackrabbits have such large ears, it would be entirely appropriate for me to respond with the teleological explanation that a

jackrabbit's big ears cool their body temperature without thereby asserting something in tension with Darwinian biology.

Essentially, what I have been describing here as the position of defenders of teleological notions in scientific explanation is the Selected Effects Theory of Functions. The Selected Effects theory is a naturalized, post-Darwinian, and explanatorily powerful analysis of functions that has established itself in philosophy of biology as the standard account of functions in the biological sciences:

> *Selected Effects Theory of Functions.*
>
> "Trait X has a proper function F if and only if X currently exists because, in the recent past, ancestors of [trait] X were successful in enhancing fitness because they performed F, leading to reproduction of genotype for trait X." (Parsons and Carlson 2008:72)[9]

The functional explanations of an antelope's stotting behavior, an eagle's wings, the human kidney, or a jackrabbit's big ears possess scientific legitimacy in virtue of providing an answer as to why these traits have been propagated and been maintained across populations in response to natural selective pressures.

How might Selected Effects functions (henceforth, SE-functions) be used to ascribe intentionality to the movements of agents? Well, we already have an example: the stotting behavior of antelope. That behavior has propagated and been preserved

[9] It is a little odd to speak of the "ancestor of [trait] X" but the idea is that a jackrabbit with the trait of tall ears has earlier jackrabbits with the trait of slightly shorter ears (but taller ones than other jackrabbits at the time) as its ancestor trait.

because it communicates to predators that they have been detected. In so doing, the antelope is able to dissuade a predator from attacking because — now having been detected — the predator tends to move on seeking easier targets. As a result, it would seem like you could say that an antelope's stotting behavior possesses intentionality or purposiveness in that it's "for scaring away predators" (i.e. possesses and serves that SE-function). So is SE-function the relevant notion for understanding the intentionality of actions? Ultimately, my answer is "no".

The first problem SE-functions face for this purpose is Davidson's swampman. Donald Davidson's swampman thought-experiment is perhaps the most serious and widely discussed counterexample thought-experiment which the notion of SE-functions has faced in the literature. Davidson's (1987) Swampman is physically identical to you but does not come about in the normal evolutionary way. Instead the swampman is a copy of you that pops into existence as a matter of random chance (e.g. due to a random collection of particles just so happening to by chance form a creature similar to a type we are familiar with).[10] In Davidson's original example, the swampman is the result of a strike of lightning but the point is just that some process other than intelligent design or natural selection could by chance give rise to a creature organized just like us. To illustrate a Davidson-esque swampman case, let's continue to rely on our already introduced example of an antelope's stotting behavior.

[10] This swampman has a seeming heart with the same function as human beings, seeming eyes with the same function as humans, seems to eat and perform other activities just as human beings do, but according to the SE-functions account would not possess any of these organs nor perform any of these actions because they lack the needed selective history to imbue those organs and movements with teleology or intentionality. If we, like Korsgaard, believe that swampman does perform actions just like normally evolved human beings do and does possess a heart and eyes just like ordinary human beings, then the SE-functions account of teleology isn't the right one to deploy for making sense of actions.

Imagine if no antelope had ever naturally evolved, but one day a swampman antelope were to by chance pop into existence just like Davidson's swampman — e.g. due to a lightning strike. Now imagine that this swampman antelope were to engage in stotting behavior just like a naturally evolved antelope does in our world. Is it not the case that such behavior on behalf of the swampman antelope would possess the same intentionality as the stotting behavior of naturally evolved antelope? And wouldn't the swampman antelope's stotting behavior possess this same intentionality despite lacking the selective history that would be required to give the stotting behavior that SE-function and intentionality in naturally evolved antelopes? If you accept that the swampman antelope's stotting behavior is also intentionally directed by the swampman antelope attempting to scare off predators just like it is for naturally evolved antelopes, then it isn't such a behavior's SE-function or facts about its origin that is responsible for its intentionality. Instead, it's something about the teleological organization of current antelopes rather than facts about the natural or intelligent selective nature of their origin that is responsible for the intentionality of the antelope's stotting behavior. Michael Thompson aims to make the same point in the following example where he takes a case of intelligent selective design rather than natural selection:

> [E]ven if the Divine Mind were to bring a certain lifeform into being 'with a view to' securing an abundance of pink fur along the shores of the Monongahela, this 'purpose' would have no effect on the inner natural teleological description of that form of life[.] (Thompson 2012:79)

In this case it isn't blind evolutionary selection that explains an organism and its traits in terms of a naturally selected-for function, but rather selection by an intelligent artificer or creator. Nonetheless the issue is the same. If a divine mind created a

certain bird with pink feathers as a means to distribute pink feathers at a certain location not by the bird living its life but simply by creating them there to die, the causal origin explanation for why the birds exist as certain living functional beings would be lacking because they still have teleological functions as self-maintaining and reproducing living systems nonetheless. That is to say, even knowing distributing pink feathers was the function which explains the creature's origin, we would look to the teleological organization of the animal itself and see different functions. The function of the bird's feathers and their pink color would seem to stem from the role they play in the life of such birds considered on their own rather than the intended purpose of the bird's creator. The function of features of the bird's life come from their place in that life rather than the purpose which led to their creation (in the intelligent designer case) or selection (in the SE-function case).[11]

[11] Alongside Michael Thompson, we find Philippa Foot giving more-or-less for the same reasons for why SE-functions is not the relevant notion of "function" for ethicists to be concerned with:

> It is imperative that the word 'function' as used here [in ethics] is not confused with its use in evolutionary biology, where, as Simon Blackburn has put it in the Oxford Dictionary of Philosophy, 'the function of a feature of an organism is frequently defined as that role it plays which has been responsible for its genetic success and evolution'. Features that are functional in this sense are what Dawkins, for instance, calls 'adaptations', when he defines an adaptation both historically and as 'approximately an attribute of an organism that is "good for something". In such contexts it is supposed to make sense to speak of the good of a species, as if a species were itself a gradually developing, one-off organism, whose life might stretch for millions of years. Perhaps the extinction of a species is imagined as a kind of death, and therefore as if it were an evil, with that which makes for its continuance thought of as 'for its good'! It is easy to confuse these technical uses of words such as 'function' and 'good' with their everyday uses, but the meanings are distinct. To say that some feature of a living thing is an adaptation is to place it in the history of a species. To say that it has a function is to say that it has a certain place in the life of the individuals that belong to that species at a certain time. (Foot 2001:32f)

Foot distinguishes between what she terms adaptations — what I take to be an alternative label for SE-functions — and true functions which seems to involve how a being's teleologically works presently rather than its evolutionary or intelligently designed origins.

The upshot is that the notion of function or purpose or teleology that is relevant to identifying the actions of agents is not the one dependent on its origin like SE-function.[12] Rather, it has something to do with the agent as they are currently constituted rather than their origin. What accounts of naturalized teleology in philosophy of biology offer such a role-in-a-system approach to teleology? It is here where I turn to Robert Cummins' Causal Role Theory of Function.

§5. The Causal Role Theory of Functions

The Selected Effects Theory of Functions — sometimes referred to as the "etiological" theory of functions — is most closely associated with the work of Larry Wright (1973, 1976), Ruth Millikan (1984, 1989a, 1989b), and Karen Neander (1991). However, there is a second naturalized theory of functions and teleology prominent in contemporary philosophy of biology and science. The Causal Role Theory of Functions in philosophy of biology — sometimes referred to as the "systems" theory of functions — is often associated with Robert Cummins and his development of the theory in "Functional Analysis" (1975:758-61) and *The Nature of Psychological Explanation* (1983:1-27). Cummins' Causal Role Theory of Functions may be either a competitor or a complement to the Selected Effects Theory of Functions. In "Functional Analysis", Cummins takes on the entrenched understanding

[12] One might object: "This does not follow. One can still say that in most cases the notion of teleology needed is the SE one, except for weird swapman cases, when we need another one." However such an objection misunderstands swampman cases. SE-effects claims that it is selective design history (either intelligent or non-designed like evolution) which is responsible for the apparent teleology of traits, behaviors, organs, etc. Swampman cases give us the same things that SE-functions claim to explain the teleology for but lacking any selective history. And in swampman cases the teleology appears to persist despite its lacking of any selective teleological origin. Thus swampman cases are not mere outlier cases, but strike at the heart of the SE-function account of teleology.

of scientific explanation as accounting for an event's occurrence or a thing's existence. Accounts of functional explanation like the etiological theory proceed with the assumption that the purpose of making functional attributions is "to explain the presence of the item (organ, mechanism, process or whatever) that is functionally characterized" (Cummins 1975:741). In response, Cummins argues that in order to correctly account for functional explanations we must recognize that functional explanation is actually a distinctive sort of explanation. And Cummins' paper had the effect of drawing philosophers' attention to an important kind of explanation being put to extensive use in many areas of science but that "was completely ignored in the philosophy of science at the time Cummins's article appeared" (Wouters 1998:102). Cummins posits a *sui generis* sort of explanation that is guided by a question of the form 'how are such and such organisms able to perform such and such activity?' rather than of the form 'why does such and such exist or exist in the form that it does?'.

Cummins notes that science isn't always interested in origins; instead, sometimes scientists are motivated to explain *how* a complex system currently functions. In these cases, the explanation researchers ultimately aim to give is one that describes how a capacity of a larger system results from the organization of components which makeup that system. While CR-functions lack scientific explanatory power with regard to questions of how or why an object or trait came to exist, CR-functions possess scientific legitimacy within the different explanatory project of accounting for how components in a larger system contribute to that system's exhibiting some more complex capacity. SE-functions, by contrast, are

explanatory in the biological sciences in the more traditional manner; i.e. because SE-functions contribute to answering the question of why a functionally characterized entity exists or exists in the form it does:

> [Cummins- or "causal role-"] functional analysis has application wherever the aim of scientific investigation is to explain the overall capacities of a complex system. One complex capacity which might be explained by functional analysis is the ability of an animal to survive and reproduce. This can be analysed into a set of simpler capacities, such as the capacity to move about, to feed, to escape predation, to mate, and so on. Each of these can in turn be analysed into even simpler capacities. In the case of feeding, the ability to ingest food, masticate it, break it down into simple nutrients, to absorb these, and so forth. These capacities in turn can be analysed into still simpler capacities, arriving eventually at such simple capacities as that of a membrane to permit diffusion of some substance. These base level capacities are directly explicable by physical laws. (Griffiths 1993:410-1)

The relevant explanandum of SE-function is why an object exists or exists in the form that it does, whereas the explanandum of Cummins- or CR-functions is how a larger system works with regards to one of that larger system's capacities.[13]

---

[13] In the view of many philosophers of biology, the Causal Role Theory of Functions has been replaced by the Selected Effects Theory of Functions as the correct theory of functions for biological science. However, an alternative view takes SE- and CR-functions to be complementary in that SE-functions are appropriate to evolutionary biology while CR-functions are appropriate to other areas. Such a complementary view of SE- and CR-functions in fact seems to be the consensus:

> Given the consensus in favor of SE function among philosophers of biology, it is surprisingly difficult to find an unequivocal rejection of Cummins's alternative. This may stem from a recognition that some areas of science (medicine, physiology, and perhaps psychology) require other kinds of function concepts. It does seem generally accepted, however, that SE function is the concept uniquely appropriate to evolutionary biology. (Amundson and Lauder 1994:444)

Those philosophers who take it to be important that we possess a unified account of function for science tend to defend the Selected Effects Theory of Functions as the only scientifically legitimate teleology. Advocates of Cummins' Causal Role Theory of Functions often endorse a pluralist view of functions in science with SE-function and CR-function taking on complementary roles. Indeed, it is not hard to see why proponents of CR-functions would be pluralist about SE- and CR-functions given the pluralism regarding scientific explanation upon which the Causal Role Theory of Functions is built.

Couch (2011) also points to Godfrey-Smith who "argues that it is wrong to think there is a unified concept of function at work in different areas. He suggests there are distinct notions of function that are appropriate to different fields. The causal role notion is appropriate in physiological investigations where researchers are concerned with understanding how the capacities of a system depend on the capacities of its components. These investigations can be undertaken independently

But there are two central sorts of objections commonly raised against the Causal Role Theory of Function: the Objection from Malfunction and the Objection from Promiscuity (or "Overbreadth" or "Permissivity"). I will address each of these objections in turn. The Objection from Malfunction is the worry that Cummins' Causal Role Theory of Functions is unable to distinguish between an object or trait's functioning correctly and its malfunctioning. A heart may be diseased and come to lack a capacity for pumping blood, but we would not say that a diseased heart has lost its function of pumping blood but rather that such a heart is "defective" or "malfunctioning" relative to the function of pumping blood it still possesses and should be fulfilling.

This may be a problem for Cummins' CR-functions because when a defective heart (or a defective anything) stops making a causal role contribution to a larger system capacity then it would seem to lose its CR-function relative to that larger system. Consequently, according to Cummins' Causal Role Theory a heart that came to lack a capacity for pumping blood couldn't be said to be malfunctioning relative to its function of pumping blood, rather we would have to say that this heart just no longer possesses the CR-function of pumping blood.

The usual way of raising the Objection from Promiscuity involves simply rehearsing a lengthy laundry list of counter-examples which illustrate this problem

---

from historical considerations about selection. Alternatively, the evolutionary notion is appropriate in areas like evolution and behavioral ecology where researchers are interested in explaining why organisms have the structures and behaviors they have. In this context, the focus is on past selection pressures in the environment, and a historical approach is appropriate. So, in Godfrey-Smith's view it is a mistake to think that we should be attempting to unify the various uses of functional language under a single account; we should rather accept pluralism. The different notions should merely be seen as reflecting the different kinds of information researchers are concerned with in different areas of investigation."

with CR-functions' looseness. The flagship in this line of attack is Millikan's (1989a) objection that while Cummins' Causal Role Theory correctly ascribes to a heart the function of pumping blood throughout the circulatory system, it mistakenly also ascribes a noise making function to the heart:

> [H]earts not only move blood through the circulatory system, but also make a thumping noise that doctors can listen to. Making a noise is an effect of the structure that can be explained in terms of the [Causal Role Theory of Functions]. But while biologists take the function of the heart to be the circulation of blood, they do not say that making thumping noises is. So the account seems too liberal since it fails to distinguish between genuine functions and mere side effects of the systems. (Couch 2011)

CR-functions are overly promiscuous (or accused to be so) in the sense that *any* sort of contribution a component makes towards *any* capacity that is possessed by *any* larger complex system constitutes a CR-function and therefore a function *tout court* of that component. As a result, we end up with hearts not only possessing pumping blood as a function but also noise-making. Relative to the capacity for pumping blood throughout the body's circulatory system, the heart's CR-function is to pump blood; but relative to the capacity for making sounds, the heart's CR-function is as a noise maker. On Cummins' analysis, there is no sense in which one of these functions would constitute a more genuine function of the heart because each causal role has explanatory power in a larger system capacity to which the heart contributes. Functional statements explain the contribution of a component to a capacity of a particular larger system, and an attribution of a CR-function is legitimate to no more and *no fewer* functions than those that have explanatory power for any capacity of a larger complex system.

In a similar case, while the proper function of feathers may be their causal role in the capacity of birds for flight, feathers also play a causal role in a bird's capacity to die from a collision with in-flight jumbo jets as well as underlying a bird's capacity for acquiring certain diseases to which they would not otherwise be susceptible. Additionally, down feathers specifically have a causal role in contributing to the capacity of the bedding made from them for contamination by molds and dust mites. Note that if our project were something like explaining the susceptibility of geese to certain diseases or the disproportionate capacity of goose feather bedding for contamination by molds and dust mites, then it would be perfectly coherent for us to frame our explanations in terms of the function of feathers in contributing to the these susceptibilities for diseases and infestations. As a consequence, when it comes to something like the function of a bird's feathers, Cummins' Causal Role Theory of Functions makes no distinction between flight, disease susceptibility, collisions with in-flight jumbo jets, or contamination by molds and dust mites. The Objection from Promiscuity may be characterized by this challenge that only flight should be considered a function of feathers.

Neither the Objection from Malfunction nor the Objection from Promiscuity pose any real threat to Cummins' notion of CR-functions. Rather, these consequences are both embraced by Cummins. If the form of scientific explanation we are engaged in is answering how a system "does what it does" then neither objection should be seen as problematic. If there is a malfunction and the system doesn't produce the relevant result, then there is no sense to us being interested in asking "how does this system produce a particular result?" It can be taken for granted that there are no

"malfunctions" when we are assigning Cummins' CR-functions because it is precisely the successful production of a certain result which we are interested in explaining when we go about assigning CR-functions as part of our explanation.

Secondly, promiscuity is merely a feature of the fact that Cummins' CR-functions aren't aimed at assigning the so-called "proper" or "real and true" or "intrinsic" function to objects but rather just the function a thing happens to play in bringing about whatever result we happen to be interested in explaining. Which is to say, once we fully embrace the entirely different explanatory project which Cummins' proposes we are engaged in when we are assigning "role-in-a-system" functions to objects, then the Objection from Malfunction and the Objection from Promiscuity dissolve as irrelevant to that explanatory goal.

Nevertheless, these two objections to Cummins' CR-functions do pose a problem for employing Cummins' proposed understanding of teleology for use to make sense of Korsgaard's view of intentional action and agency. The stotting behavior of an antelope upon noticing a stalking predator is an action the antelope performs in order to signal to that predator that he's been detected and hopefully thereby avoid being attacked. That is the function of the behavior and determines the behavior's intentionality — i.e. what the antelope is trying to do. But if we adopt Cummins' CR-functions there is no special sense in which this is uniquely the function or purpose of the antelope's stotting. The behavior has perhaps infinitely many functions depending on what we are interested in explaining and how antelope avoid predators or survive is just one among many with nothing special about it. This is the problem with promiscuity.

Furthermore, suppose that we were able to say that the antelope's stotting behavior had its function and thereby intentional content fixed by — say — the behavior's role in the antelope's survival. Even then, if the stotting fails to dissuade the predator and he attacks, kills, and eats the antelope then neither the stotting nor the antelope has failed in such a case because then the stotting wouldn't have had the function of scaring away a predator. Since Korsgaard allows that an animal's behaviors can have a certain intentionality but fail, the account of teleology her view employs must allow for the possibility of malfunction. Afterall, consider what would result if Korsgaard didn't have a notion of function which allowed for malfunction. An antelope stottles but fails to do so loudly enough to scare away the predator who is stalking it, so now the behavior no longer has the Cummins' CR-function of scaring away predators because it in fact doesn't. On this account the antelope hasn't failed to do what it intended to do, but rather it never performed an action with intentionality at all since intentionality requires actually performing a certain role in a particular system outcome. This will not do.

This concludes our tour of the naturalized accounts of teleology that have been proposed within philosophy of biology to make sense of the notion of teleological explanation post-Darwin. In the next section I will take stock of where we are at and why none of these accounts seem to fit the needs of the teleological notion the account of agency and the account of intentionality I laid out at the beginning of this paper require.

§6. Taking Stock

The problems with the traditional SE-function and CR-function account for the view of agency and intentionality has been raised by previous authors. Let's start with the Selected Effects account of function. What Davidson's swamp man case is meant to show is that an agent without a selective origin via an intelligent designer or evolution still intuitively displays intentional and teleological organized behaviors. That is, even a swampman duplicate of you randomly created by a lightning strike in a swamp would be correctly ascribed a heart that pumps blood in order to circulate oxygen and keep you alive despite the absence of any selective history and that when your duplicate pursues food that he's doing so to satiate his hunger despite the lack of any historical selection in bringing about such behavior. Davidson "bites the bullet" here and accepts the idea that the swampman heart lacks the function of pumping blood because the swampman's heart lacks a selective history directed toward the function of pumping blood – but this isn't a bullet we should bite:

> Historical theories of functions face rather different objections, but the most powerful objection is that we do seem to attribute functions without any commitment as to the history of the entity to which we attribute a function. This is vividly illustrated by Donald Davidson's example of a 'swampman' who is a molecule for molecule duplicate of Davidson who just 'pops into existence' in a swamp (Davidson 1987). Davidson bites the bullet in this case, but many would find it implausible to suppose that Swamp-Davidson's heart does not in any important sense have the function of pumping blood. (Ridge et al. 2011:370)

This isn't to say the SE-function account is incorrect for the purposes of philosophy of biology, it's that the SE-function account isn't the one needed by a theory of agency and intentionality; An SE-function account might be what's needed by evolutionary biology without being the correct account for understanding agency and

intentionality. The problem for the account of agency I've provided is that ascribing intentionality to a movement intuitively depends upon the *current* teleological arrangement of a thing rather than the historical role that selection has played in its origin – something which is exemplified by Davidson's swampman case.

Robert Cummin's ahistorical-system, causal role account of function does not depend upon selective origination but instead on a component's causal-role in a larger system's outcome.[14] This seems like a better candidate for the conception of teleology needed for the account of agency I've advocated, however, the CR-function view has the problem of being unable to account for malfunctions and the problem of promiscuity. However, there is an Aristotelian conception of life that may ameliorate this problem. Once we consider Aristotle's account of life — i.e. a self-reflexive system — then we might be able to employ the Cummins' CR-function notion in a way that avoids the dual problems of malfunctions and promiscuity.

The crucial idea is that unlike ascribing functions only to components insofar as they play a role in a system's capacity to produce a certain outcome, there is also an ascription to the whole system and its overall outcome. Specifically, the CR-function that is ascribed to the overall system is the system supporting its parts in promoting the system. In order to have a system where the outcome itself also has a purpose we require a teleologically organized system which is circular or "self-reflexive".

[14] At one point Ridge et al. even suggests Korsgaard's account of function seems to be equivalent to Cummins' CR-function: "Korsgaard's conception of function [...] is closer to the [Cummins'] systems capacity conception than the [selected effects] historical conception. [...] In fact, this sounds like the 'systems capacity' conception of function. [...] [In one way of reading Korsgaard] we have a view that is more or less identical to Cummins's classic systems capacity conception of function." (Ridge et al. 2011:370-2). However, later Ridge points out that Cummins' account could not serve her purposes.

This idea of a "self-reflexive system" or a "self-maintaining system" – i.e. a living system — is worth taking slowly. It is a system which isn't just teleologically organized toward producing a certain outcome (even artifacts like alarm clocks are like that), but rather the outcome that its parts are organized towards must be to self-maintain as a system which produces itself. In the next section, I will explain and develop this idea. The notion of a self-reflexive system — i.e. a living system — that provides us with a form of Cummins' CR-functions which avoids the problems this view of teleology traditionally faces.

§7. Self-Reflexive Systems and the Problem of Malfunction

Korsgaard introduces her notion of function as "how something does what it does". The idea here is that anything that as a matter of its essence "does anything" (judged so by whatever standard you like) and also has a matter of its essence "how it does it", possesses a function. This is essentially the Cummins' CR-role notion of function and such has been pointed out by other commentators:

> If we understand 'what something does' as the capacity of a containing system, and 'how it does what it does' in terms of the contribution of some simpler contained parts to that broader capacity, then we have a view that is more or less identical to Cummins's classic systems capacity conception of function. (Ridge et al. 2011:372)

On Korsgaard's account, a function is how something does what it does. What it does seems to simply be whatever causal outcome a thing — i.e. a system — produces, and how it does it seems to simply be a Cummins' CR-role functional analysis of how the parts contribute to this particular outcome. However, Korsgaard quickly shifts from a direct reliance of CR-function when she introduces Aristotle's notion of life:

> [A]ccording to Aristotle, a living thing does have a definite purpose, in the sense of a 'what it does'. That purpose is to keep its own form, its own manner of functioning, in existence. It does this in two ways: first, through the continuous self-rebuilding activities of nutrition, which maintain its form in a spatio-temporally continuous stream of matter, and second, through reproduction, by which it imposes its form on individually distinct entities. This is not a controversial metaphysical thesis about what living things are for, but rather a definition of 'living'. If a thing has a form that is self-maintaining in these basic ways, then it counts as 'living'. So far as this goes, there is nothing objectionable about Aristotle's teleology. The appropriateness of teleological explanations need not have anything to do with claims about how or why the object whose parts and activities we seek to explain came into existence. Teleological explanation may be appropriate to an object simply because it has a self-maintaining form. We seek such explanations when we ask what contribution its arrangements or parts make to its self-maintenance. (CA 141)

With a system where CR-functions can be assigned in a circular, self-reflexive manner — i.e. "living systems" in the Aristotelian sense — this has an important effect on how we understand such a system's CR-function. According to Cummins' CR-functions, the overall systems have no function.[15] There is a pragmatic, social, or otherwise practical interest in knowing how a particular system is able to produce a certain non-teleological, causal outcome and by assigning the parts of a system CR-functions we come to understand how the system comes to produce that outcome. However, if the system itself is "circular" or "self-reflexive" then that produces an important change which leads to the system itself being able to be assigned a CR-function in addition to its component parts. Let me explain.

In order to introduce the idea of a self-reflexive system — a living system — I want to consider the most basic form of life I can hypothesize. For this reason, I want to think about a very basic first living creature existing back in the primordial soup.

[15] Except insofar as a component in yet some further overall system — but this nesting of systems must halt somewhere with ultimately an overall system lacking any function.

Imagine an environment of some rich primordial soup as well as some form of structure that absorbs building materials (e.g. electrons, carbon, oxygen) from that primordial soup from which it assembles itself into that very structure that absorbs building materials from this environment in order to assemble itself. That is, imagine a simple self-reflexive structure engaging in its self-reflexive activity in the primordial soup. To be more specific, let me introduce the idea of my Primordial Organism and the environment in which it lives:

> **Primordial Soup.** Suppose the primordial soup in which this simple lifeform exists consists of two basic materials M1 and M2. Further suppose that the soup also contains a more complex material, CM1, made out of two additional basic materials M3 and M4. Basic materials M3 and M4 do not naturally occur in the primordial soup but are only present in the combined form CM1.
>
> **Primordial Organism.** The organism we are to imagine has a double outer shell with a single flap-like opening on one side. Immediately inside of the flap-like opening is an antechamber to the interior of the cell. Due to the lower pressure that exists within this first cavity within the cell, material from the primordial soup is pulled into this antechamber through the organism's outer shell's flap-like opening. Once inside, the materials taken from the primordial soup have less room to maneuver, this causes the materials in the antechamber to collide with one another and consequently break up the more complex material CM1 into its constituent materials. This process creates M3 and M4

> within the organism's antechamber. The basic material M1 is electromagnetically strongly repulsed by the newly created basic material M4 in a way it was not repulsed by CM1, causing most of the M1 material within the antechamber to rapidly decompress out of the antechamber. However, not all of the M1 material exits the organism. Some of the material filters through small openings in the organism's inner shell so that it lodges in the space between the inner and outer shell. Over time this void between the organism's inner and outer shell becomes filled with M1 material and forms the organism's new, replacement outer shell for when its original outer-shell becomes more and more damaged overtime by impacts from debris in the primordial soup and eventually dissolves away. Next the newly created material M3 within the antechamber moves into a secondary chamber within the organism due to… and so on like this until we've described a complete process whereby this primordial organism self-repairs, maintains, and so rebuilds itself by drawing from materials in its environment and processing them into becoming the organism's replacement parts.

Here is a system whose parts and behaviors — its antechamber, its double inner and outer-shell with porous openings on the inner-shell for repelled M1 material to enter into the void between, the sucking in of material from the primordial soup, the breaking apart of complex material CM1, etc. — can all be assigned a function or purpose with respect to how each contributes to the way the system re-builds and maintains itself. However, there's more to it in this case. There is more to it in this

case because this system is self-reflexive. It is not just that the parts that get attributed functions or purposes in our explanation, but additionally the system itself also comes to be assigned a function or purpose. The parts have a function in contributing to the whole and that is how we conceptualize the parts, the whole has a function in contributing to the parts and that is how we conceptualize the whole.

It is important to realize how this self-reflexive structure makes such a transformative difference. Suppose we were to ask why our primordial organism expels M1 material into the void between its inner and outer-shell. There is one answer that is causal and does not involve functions or teleology. However, notice what occurs if we attempt to give a teleological explanation of why this happens in terms of the role of the organism expelling M1 material into the void between its inner and outer-shell towards the larger system of which it is a part. If the system loops back on itself, then each part as well as the entire system could be ascribed a function. In such a case there would be a complete teleological explanation. I could explain why the system exists teleologically without need for appeal to a causal explanation. *I could want a mechanistic causal explanation*, and the teleological one won't give me what I want then — but a mechanistic causal explanation won't give me a teleological explanation either. At the beginning of this paper I introduced the anti-causalist view of agency wherein teleological and causal explanations of actions are *sui generis* and independent. We now have come to return to that view. A self-reflexive system can have its parts explained in terms of their CR-function contributing to an outcome produced by the whole, and the whole can be ascribed and explained in terms of the CR-function it possesses promoting its own parts. And so a

teleological explanation of why an agent does this-or-that or exists might be circular, but it can be complete.

Here's another way of raising the same issue: "But why is possessing a self-reflexive function so important to agency? Why can't artifacts like watches or alarm clocks be agents if plants can be agents?" Components of a teleologically organized object cannot exist separate from the whole. A slab of sheetrock or plaster *can* exist separately from a house, but a wall *cannot*. A wall "divide[s] one room from another" or "help[s] to hold up a roof." (Korsgaard SC §2.2.2). It is only when appropriately integrated into a house that a slab of sheetrock or plaster becomes a wall. There is no possible world that contains walls but no houses. But notice now how the same dependency reoccurs on the level of houses. A house, say, provides shelter to animals. Just as there cannot be walls without houses, there cannot be houses without tool-using animals (in this case, the "tool" being a house). This chain of teleo-functional objects must come to an end with a teleo-functional object with a self-reflexive end or else no teleo-functional object exists.

Notice that in this chain from walls to houses to animals that once we reach tool-using animals the dependency does not reoccur. This is because tool-using animals have self-reflexive forms. A living thing's heart functions to pump blood in support of the creature's overall end, and the creature's overall end of self-maintenance promotes the functioning of, e.g., the creature's heart. Consequently, when it comes to agents with self-reflexive forms, their agency is not dependent on being a part of a larger agent in the way the functioning of walls, houses, alarm clocks, and hearts do.

The idea is then that houses do not function to provide shelter, but rather are a living agent's way of sheltering itself. Similarly, the alarm clock doesn't successfully wake you up by ringing at the right time or fail to wake you up by malfunctioning, it's you that succeeds or fails to wake yourself up using the alarm clock. It is not really a house or alarm clock that ever succeeds or fails, but the living agent who succeeds or fails in its use of such artifacts. If you'll remember I have argued that the form of intentionality or "purpose" necessary for a movement to be an action is one that renders coherent a judgment that the agent has succeeded or failed at what it's trying to do. Here I am arguing that really it is only living systems — i.e. self-reflexively organized teleological systems — which can be judged this way since tools and machines like alarm clocks only succeed or fail relative to us whereas inanimate objects like rocks and the ocean never succeed or fail. Only a self-reflexive teleological system can be judged to succeed or fail according to the system's own purpose (i.e. to self-maintain its parts in those part's own functions to self-maintain the whole).

So far this doesn't make self-reflexive systems unique. It is a salient distinguishing feature of living systems like this one that the overall resulting system is itself assigned a role in by our conceptual dividing up the world of producing itself, however it isn't clear what makes this ontologically special or why when science is "carving nature at its joints" we need to specify self-maintaining systems as a distinct kind. After all this doesn't enable us to causally explain anything we couldn't before

nor does it enable us to make predictions that we couldn't have made without positing self-reflexive systems as a special sort of kind.[16]

However, we have philosophical or ethical grounds for recognizing this as an importantly distinctive kind in the world because of how — for example — it enables us to make sense of the notion of "natural goodness". Korsgaard has argued that having a circular or self-reflexive teleological system like with life is necessary for us to be able to make sense of the notion of natural and/or final goodness. If we speak of what's good for alarm clocks and hearts, we merely mean what's good for such things to function well as the type of tool which is useful for us — it's "the sense in which 'good for you' [i.e. the alarm clock or heart] means something like 'enables you to function well [as the thing you are].'" (Korsgaard 2015:146). However, there is no sense that an alarm clock functioning well is good *for* the alarm clock — although its functioning well might be good (or bad) for us and qualify as well functioning for the alarm clock qua alarm clock, being good for us or good qua alarm clock isn't good for the alarm clock itself. For creatures with self-reflexive ends (i.e. an end of preserving oneself as a system which self-preserves in the particular way one does), we can talk about things being good for it in that it promotes its functioning, and then its well functioning in turn promotes its performance of those very activities.

[16] One might ask, "Is self-maintaining the same thing as self-reflexive?" The answer is "no". A system that is self-reflexive is one that can be conceptualized as a teleologically organized system specifically directed at preserving its self-maintaining component parts and processes to add-up to accomplishing the ultimate goal of preserving itself as self-maintaining itself via these component parts and processes. Which is to say, there is a certain manner of self-maintaining — as a spider does it by spinning a web and catching prey — which the wholistic system (including the spinning of a web to catch prey) is just itself a part in maintaining the system of self preserving a system that self-preserves in just this way. I take it that many systems can be self-maintaining in ways that preserve their existence in limited ways without the whole system solely being ultimately understood and conceptualized as singularly directed at self-maintenance.

Korsgaard labels the well functioning of a living creature — i.e. a teleo-functional kind with a self-reflexive end — "health". And she describes the situation thusly:

> [I]t is good for people to enjoy themselves sometimes, contemplate natural beauty, go to museums to appreciate art and learn something about the world they live in, so that they achieve a state we call 'health,' which in turn promotes their ability to enjoy themselves sometimes, contemplate natural beauty, and go to museums to appreciate art or learn something about the world they live in. What sort of a merry-go-round, you might ask, are we on here? (Korsgaard 2015:143)

Once we have self-reflexive teleologically organized systems, then, we come to also have a notion of (non-prescriptive) natural goodness.[17]

That is, I can speak of what's good or bad for, e.g., a plant without implying that I or anyone else or even the plant has a reason to act towards what's good for the plant.[18] Nevertheless, once we have introduced self-reflexive systems, there is a notion of "a thing's good" which can be spoken about which couldn't be before. Another illustration Korsgaard offers is a horse compared to a pruning knife:

> Aristotle himself uses the example of a horse, and says that the virtue of the horse ''makes a horse both good in itself and good at running and at carrying its rider and at awaiting the attack of enemy'' (NE 2.6 1106a19). But it is not obvious that a horse achieves its own good in being ''a good horse'' if what that means is a horse good for human military purposes. Might not a skittish unmanageable horse win for itself a fine free horse-life away from the dangers of warfare? One of Plato's examples is a pruning knife (R 353a), but it would

[17] If it isn't clear at this point, I should make clear that even bacteria qualify as agents on this view. It's possible that ecosystems as well as computer viruses also count. I am not sure about those latter pair, but the standard of "life" and "agent" here is very broad. Human agents do possess a very narrow and special sub-set form of agency, but the broad concept of agency is quite broad on my account. In her most recent work Korsgaard offers new reasons to restrict the notion of agency just to those agents who can experience pleasure or pain, but that restriction is somewhat contrived and even if we accepted it that wouldn't disprove the way I use the term "agent" but rather reduce to a terminological dispute about the most useful manner for how narrow we restrict our application of the term "agent". I take how loosely or narrowly we use the label "agent" to merely be terminological.

[18] Korsgaard explains at one point: "[E]ven if a plant does not have a final good in the sense of something to aim at, the good of a plant is "final" in a different sense. The good of a plant is "final" in the sense that the explanation of why things are good-for it ends with the plant." (FC §11.4.4).

> be absurd to infer that a good pruning knife achieves the good for a pruning knife. (CA 131)

We might speak of what makes a good horse given our interests just as we might speak of what makes a good pruning knife given our interests. However, in the case of the horse but not the pruning knife, there is also sense in talking about what's good for the horse itself and not relative to any other being's interests. And so the notion of self-reflexive teleological systems is the basis for our notion of "good" in some sense of good.

It is at this juncture virtue ethical philosophers like Phillipa Foot seem to become convinced our work is over. Foot asserts that a component's function can be identified with the causal role contribution it makes in the life cycle of the organism to which it belongs. Rustling is not a part of the function of tree leaves because while this may play a causal role in some larger complex system, it does not play a causal role in which it contributes to the capacity of a tree to complete its life cycle. Foot doesn't use the concept of self-reflexive teleological systems, but rather points to the work of Michael Thomson for his definition of life in terms of "natural-historical judgements" or "Aristotle categoricals" — which I think can be read as actually speaking of self-reflexive teleological systems when viewed in the terms of CR-functions. Here's how Thomson describes the idea:

> In learning of the various cellular processes unearthed and described in biochemistry — photosynthesis, for example, or the Krebs cycle, or the replication of DNA — one is inclined to think, It's all getting boiled down to chemistry and physics, isn't it?, and in some sense of 'boiling down' this is of course true and very desirable. But it is interesting that if the only categories we have to apply are those of chemistry and physics, there is an obvious sense in which no such succession of goings-on will add up to a single process. In a description of photosynthesis, for example, we read of one chemical process — one process-in-the-sense-of- chemistry, one "reaction" — followed by

> another, and then another. Having read along a bit with mounting enthusiasm, we can ask: "And what happens next?" If we are stuck with chemical and physical categories, the only answer will be: "Well, it depends on whether an H-bomb goes off, or the temperature plummets toward absolute zero, or it all falls into a vat of sulfuric acid . . ." That a certain enzyme will appear and split the latest chemical product into two is just one among many possibilities. Physics and chemistry, adequately developed, can tell you what happens in any of these circumstances—in any circumstance—but it seems that they cannot attach any sense to a question "What happens next?" sans phrase. The biochemical treatise thus appears to make implicit play with a special determination of the abstract conception of a process, one distinct from any expressed in physics or in chemistry proper. [...] [I]t is not just that 'the rose and maple are subjects of processes of their own': they are also subjects of a special type or category of process—'biological' processes, if you like, or 'life-processes'. [...] [S]uch processes are after all what goes on as life goes on [...] events as bound together in a life-process. (Thompson 2012:41-3)

Elsewhere he points out without diagnosing why that living organisms are uniquely susceptible to "natural-historical judgements" or "Aristotle categoricals" — I take the explanation to be the fact that living organisms are self-reflexively teleologically organized systems subject to circular CR-functional explanations:

> A life-form or species (in the broad sense) is anything that is, or could be, immediately designated by a life-form-concept or a life-form-word. [...] Or, equivalently, an organism is the object of any possible judgment, this S is F, to which some system of natural-historical judgments, the S is G, H, etc., might correspond. If an intellect loses the capacity for the latter sort of 'synthesis' it must also lose the former, and with it, I think, the capacity to experience things as alive. (Thompson 2012:77-8)

For Foot, the existence of kinds which can be conceptualized as self-reflexive systems means that — as a scientific, theoretical matter — they exist and a complete description of the world must include them:

> Some intelligent Martians who themselves did not think in terms of goodness and badness might (even if landing in the rain forest and knowing nothing of humans) realize that the plants and animals on earth could be described in propositions with a special logical form, and come themselves to talk about the newly met living things as we do. They would rightly see the existence of this different order of things in the world as an extremely interesting

> ontological fact, allowing them to invent and deploy a range of concepts that they did not have before. (Foot 2003:36)

I will agree with Foot that once we come upon systems which can be conceptualized as self-reflexive in this way, we become able to invent and deploy a range of concepts we did not have before. However, Foot is so focused on the question of whether she could deploy these concepts she fails to stop and consider the question of whether she should. Perhaps we can conceptualize the structure sitting in the primordial soup teleologically as a self-reflexive organism and this would enable us to speak of natural goodness and what is good for it, but we so far lack any justification for conceptualizing the world in this way.

Korsgaard differs from Foot on why we must include these self-reflexive systems in our conception of the world. Unlike Foot, Korsgaard does not seem to think that a fully scientific description of the world requires that we recognize self-maintaining systems as distinct kinds. Korsgaard doesn't see anything about our causal-scientific description of the world to require us to recognize self-reflexive creatures as a supervening kind. Instead, there is a practical necessity which Korsgaard believes imposes this way of conceptualizing the world upon us. Korsgaard argues that seeing oneself as an agent and the world in general populated by other agents is a way of conceptualizing the world that's necessarily built into the first person point of view of acting:[19]

> The idea that teleological thinking is inherent in our powers of conceptualization is a development of a point that is implicit in what I have already said. A teleological conception of the world is essential to our functioning as agents. (SC §2.3.2)

[19] If we adopt an interventionist conception of causation, then the very notion of cause and effect depends upon the idea of agency and so requires this point of view.

Why does Korsgaard make this ambitious claim? I will reserve what I to say about this for later work where I have sufficient space to turn from speaking about agency in general toward more specifically give an account of human agency and free will. But here is a brief preview of the difference between Foot and Korsgaard on this topic: Consider the issue of determinism and the various ways in which we might interpret the threat that determinism can pose to our agency.

One manner of viewing the threat of determinism is that any action I take is, in fact, the result of a causal chain extending back not just before my decision now or even before I was born but all the way back to the very initial state of the universe. How then can my actions be attributable to me rather than the initial state of the universe? That's one manner in which we might think about the threat of determinism.

But there's another manner of understanding the threat of determinism. When I raise my arm, this does not just depend upon a causal chain extending back to the initial conditions of the universe but also upon everything else that's going on in the universe. Had there been a comet about to strike the Earth, then I wouldn't raise my hand — I'd be obliterated. Had there not been the planetary factors which kept the oxygen I breath contained on the Earth's surface and in the room I occupy, then I wouldn't have raised my hand because I'd be too busy suffocating and dying. Had the causal chain responsible for the continued homeostasis of the Sun not been present — e.g. had the Sun suddenly gone out or gone supernova — then I wouldn't raise my hand because I'd either be burned to a crisp or a frozen corpse. Had the plaque in my arteries broken free and caused me to have a heart attack, then I wouldn't raise my

hand because I would be dying. In fact, anything I do doesn't just depend on a causal chain that extends all the way back to the initial state of the universe, *it also depends on absolutely everything in the universe up until the moment before my hand goes up*. Allow me to explain.

Korsgaard points out that once we go looking for mechanical causes to explain what happens, then this will not only trace ever backwards but also come to involve the entire previous state of the universe. Viewed deterministically, then, it's not just that every causality is determined by an event causal chain traceable back to the initial state of the universe, it's that there is in fact only one causality that could possibly exist: the state of the entire universe in the previous moment. And there is only one possible effect that can occur: the state of the entire universe in the present moment. Viewed mechanistically, there are no individual causes populating the universe but only the state of the universe at one time causing the state of the universe at another time:

> From the purely mechanistic point of view, the identification of a particular object or even a particular event as the cause of another is artificial, a piece of shorthand, a sort of conception of thumb, if I may put it that Way. [...] Now perhaps some people suppose that as long as you conceive the knife merely as the cause of the cutting, rather than as for the purpose of cutting, you are not conceiving of the world teleologically. The view that the knife is the cause of the cutting is mechanistic. But is it? In the purest form of the mechanistic view, the knife is not the cause of the cutting. [...] Assuming something like determinism is true at the level of middle-sized objects, the cause of the cutting is the state of the world a nanosecond ago determining the state of the world now. Why then do we say that the knife, rather than the state of the world a nanosecond ago, is the cause of the cutting? (SC §2.3.2)

For this reason, Korsgaard argues that what's necessary in order to see ourselves as agents (i.e. individual casualties), is that we must have another way of conceptualizing the world other than mechanistically. Specifically, we must be able to

view the world as made up of teleologically conceptualized individual casualties rather than mechanistic causation. For if we conceptualize the world mechanistically instead of teleologically, there is no such thing as an agent or any individual causality producing effects attributable to it — only the whole state of the universe at one time causing the whole state of the universe at another. Mechanical causation cannot identify a cause of an effect except by citing the entire previous state of the world. Viewed non-teleologically, not even electrons would have effects. It is the whole state of the universe that produces the whole state of the universe at the next moment, when the world is thought about mechanistically rather than teleologically.

There are then, Korsgaard argues, two possible pictures of the world. In one, cause is mechanistic and always simply involves the state of the universe at one time causing the state of the universe at the next. In the other, cause is teleological and involves individual things causing effects on other things. It is this second picture of the world that is necessary to our conception of ourselves as agents — i.e. as individual causalities.[20]

---

[20] Now, this view of Korsgaard's is — to say the least — metaphysically ambitious and to fully examine or explain would take us far afield. However, there is one connection between Korsgaard's argument here and my earlier discussion of Cummins' CR-functions that I think is well worth bringing out before moving on. Cummins' contends that his form of scientific explanation is distinct from the causal, origin-focused form of explanation which — by contrast — the notion of an SE-function offers us. Cummins' argues that we can find a different form of scientific explanatory satisfaction in taking some mechanical-causal outcome of a system and then thinking teleologically about how that system produces such a mechanical-causal outcome. But — upon reflection — this is a strangely bi-furcated form of explanation. It starts with a mechanical-causal outcome of the overall system and then explains as the teleological product of the system's subcomponents. It neither assigns that final outcome of the whole system a teleological CR-function nor does it attempt to explain it teleologically. This, I suggest, is a symptom of the fact that Cummins has not taken his insight far enough — at least if Korsgaard's views about the divide between the mechanistic and teleological view of the world is correct.

At least with regards to self-reflexive systems, the teleological explanatory frame of mind can be taken to its completion. The causal explanation circles back on itself such that the whole system has a CR-function which depends on how the system promotes the functioning of its parts, and the parts have CR-functions by how they promote the functioning of the overall system. The explanation is circular, but it is complete. Cummins attempts to invoke the teleological mode of thinking to offer

§8. Two Standpoints

At the beginning I introduced the anti-causalist theory of action as one that treated teleological and causal explanations as wholly distinct. To explain causally why a motion happens as a matter of mechanistic efficient cause is a wholly different project from giving a teleological explanation of why the action occurs in terms of its intentionality. Now we have returned to that thesis. There are self-reflexive systems which can be conceptualized teleologically which allows us to think about and explain them wholly teleologically. To conceptualize such beings this way serves no role in our causal picture of the world and need not appeal to it. In effect, a self-reflexive system can be thought of as a universe unto itself and explained in teleological terms unlike everything else which can only be explained by appeal to a causal story.

There are these parts of the world which can be conceptualized teleologically rather than causally. There is nothing which makes them special scientifically in the sense that we require them for building our causal story of the world, but they are strikingly vulnerable to this other form of thinking and conceptualizing the world. We can give an explanation of these things and why they exist which is teleological.

This doesn't violate naturalism — at least not in one sense of understanding such a view. We can give a causal explanation for how systems which can be

---

explanations without fully embracing this teleological mode of cognition. Cummins' explanations always proceed from the assumption of a system having a merely mechanical-causal outcome which we then adopt the teleological mode of thinking to explain. I take Korsgaard's view to be at odds with Cummins' proposed form of explanation. There is no legitimacy to the sort of explanation being offered by Cummins, because to truly adopt the teleological mode of explanation is to seek a fully self-sufficient explanation within that mode itself. Cummins-style explanations give the appearance of teleological explanation, but it stops short of seeking a full teleological explanation (which will require a self-reflexive system at some point).

conceptualized as self-reflexive teleological systems causally arose. They are not magical in the sense that their teleology cannot be fit into the naturalistic causal picture of the world, but to conceive of them teleologically is not to stand in need of such a causal explanation for why they exist. I can give a fully teleological explanation for them.

This, I propose, offers us a possible manner of understanding agency along the lines of traditional anti-causalism. Agents are self-reflexive teleological systems. The explanations of their actions is by placing their movements in these systems and it is sui generis from causally explaining their movements. I've argued here that the circularity of self-reflexiveness is important to opening up this alternative way of conceptualizing matters.

Classic Cummins' CR-functions explain how a system produces a result, not why the result occurs or why the components of the system behave as they do. So unless we modify our way of thinking about CR-functions when it comes specifically to self-reflexive systems, then there is no way to answer a "Why?" question with Cummins' CR-functions as classically conceived without them "topping" out as purposeless causal outputs of the whole system.

This leaves us with SE-functions, which at first seem promising. SE-functions aren't teleologically complete. The function of a bird's wing is flight, but the function of flight doesn't explain by itself why birds have wings unless we add to our explanation a causal process which generates and then selects for traits that fulfill the purpose of flight. The function of wings may play a role in the causal explanation of why birds come to have wings via the process of natural selection, but the teleological

explanation that wings are for flight does not explain why birds have wings itself. If I ask what the function of birds' wings are, the teleological answer is "to fly". Why does that explain its existence? Because of a causal process that selects for traits which increase odds of procreation.

So classically conceived of CR-functions don't offer us a parallel, *sui generis* way of answering 'why' questions but instead 'how' questions. The SE-function account does explain where and why SE-functions are ascribed, but unfortunately the explanation is causal. So neither of these accounts of teleological function result in a separate teleological form of explanation separate and autonomous from causal explanations. But when Cummins' CR-functions are applied to self-reflexive systems, we do get a fully distinct and teleological way of answering why questions differently from causally. Which is to say CR-explanations can be complete — *at least in the special case of self-reflexively teleologically organized systems.*

§9. Conclusion

I began in §§2-3 with the debate between the anti-causalist and the causalist in philosophy of action over whether reasons are causes. I then turned to Christine Korsgaard and her argument concerning why the causalist view of action cannot be correct because of how it gets the intentionality of actions wrong. Korsgaard argues that actions gain their intentionality from the teleological structure of the agent performing that action. This then raises difficult questions regarding how to understand the sort of teleology that Korsgaard must implicitly be relying upon here and so I spend §§4-7 examining the various accounts of teleology which might serve her purposes. In particular, §4 I considered the Selected Effects view of teleology, in

§5 I considered the Causal Role view of teleology before taking stock of the failures of each of these accounts to serve the needs of Korsgaard's notion of agency in §6. After finding all three of these accounts of teleology lacking, I went on to propose in §§7-8 that if we use Aristotle's notion of "life" as a self-reflexive system that we might be able to modify two of these previous accounts of teleology in order to make sense of our notions of agency and action.

In the end I hope to have shown how by taking Cummins' Causal Role Theory of Functions and applying it to systems which are internally organized toward self-maintenance and preservation — a.k.a. are "alive" in the Aristotelian sense — then we can come to a view of agency where agents are part of a teleological realm that cannot be reduced to the lower level of non-purposeful mechanistic causality upon which the realm of agents supervene. In later work I intend on building upon this account of agency in general — wherein all "life" in the Aristotelian sense qualify as agents — and develop a more specific account of human agency and the manner in which human agency is uniquely a form of free agency.

Paper 2: Human Life, Teleology, and Libertarian Free Agency

§1. Introduction

Actions are always performed "on purpose". A lioness follows the scent of her prey *in order to* hunt and feed, I raise my arm *in order to* get the attention of a friend, an agent ducks *in order to* avoid being hit by an object flying in their direction, and so on. The words "in order to" in each of these descriptions single out the purpose or reason behind a particular agent's action that could be given as a teleological explanation for why they did what they did.[21] But how can this be true? How can actions — or anything for that matter — be explained by giving the purpose or reason for a movement rather than its prior efficient cause? On the received view within philosophy of action, this actually isn't possible. Afterall, reasons and purposes don't make things happen, causes do.[22]

---

[21] Throughout I use the terms "teleological explanation", "intentional explanation", and "rational explanation" interchangeably. In the literature in philosophy of action these phrases are also typically used interchangeably. Additionally, I follow Davidson here in using the term "rationalizes" not in its now common derogatory sense where someone confabulates reasons to make an irrational action seem as-if it had been performed for good reason. Instead, I follow Davidson's use of the term "rationalizes" here to just mean "provide rational justification" or "make a movement rationally explicable by citing justifying reasons" or "rationally intelligible".

[22] In certain rare circumstances it may be more appropriate to explain an agent's waving of their hand above their head by tracing the causal chain from the agent's neural activity down to the particular muscle contractions which ultimately led the agent's arm to wave. For example, this causal sort of explanation may be appropriate in the context of a psychology lab attempting to understand how the neuro-physical workings of the human brain serve to regulate human movement. However, this is a very specialized context and one even a working neuroscientist rarely finds themselves in. Usually when someone asks why an agent might have waved or performed any action, the sort of explanation we are after is a teleological one rather than the causal one. That is, we want to know the reason or purpose for the action rather than its efficient cause in the agent's brain or arm. We want to know if the agent was waving her arm "in order to get her friend's attention" or "in order to stretch" or for some other reason rather than wanting to know how the agent's waving causally came about from the inner causal workings within the agent's physical body. The story is a bit more complicated than this because of the relevance of causal explanations to irrational actions in addition to the neuroscientist's lab, but I'll as a means of introduction to the topic I'm ignoring that further complication here.

Since Donald Davidson, the standard view within philosophy of action has been that only causes can explain. Consequently, if *the reason for* or *purpose of* an action possesses any explanatory power, it must only be because an agent's reason or purpose is part of the causal explanation for that action. In slogan form, this view maintains that "reasons are causes". In such a "causalist" view, an action is explained and brought about by causes just like any occurrence — albeit a particular sort of cause. The causalist's claim is that when we explain an agent waving her arm by citing her purpose or reason for doing so as an attempt to get the attention of a friend in the distance, we are actually just attributing to such an agent, e.g., the mental states of seeing a friend in the distance and a desire to get that friend's attention which together combine to both (1) cause the agent to raise and wave her arm, as well as also (2) provides us with an intelligible purpose or rationale for why the agent would be waving her arm. The fact that actions are subject to teleological explanations therefore is simply a result of the sort of cause that actions require to constitute an action. Specifically, actions are movements with mental state causes which pull double-duty both casually explaining the movement as well as enabling us to see a purpose or rationale for why the action was taken.

I will begin in this paper by exploring one of the key reasons we must reject the traditional causalist view of agency. In other work I've argued against the traditional causalist view because of the long-standing "problem of deviant causal chains"; in this paper I raise what's known as the "problem of the disappearing agent". Both provide us with sufficient reason for abandoning the standard causalism toward actions. Actions do not inherit their purposiveness from their mental state

causes; instead — as I argue — an action's reason or purpose stems from how a movement or effect is embedded in a teleologically organized agent's way of life.

Central to the alternative account of agency I defend in this paper is the notion of "life" as a self-reflexively teleologically organized system. Each form of life isn't just teleologically organized toward producing a certain outcome (even artifacts like alarm clocks are like that), but rather the outcome that a living organism's parts are organized towards is always to self-maintain as a system which serves to produce itself. Such a non-standard teleological and non-causalist theory of agency has its roots in the work of Aristotle, Wittgenstein, and Anscombe, but in particular I aim to build upon the version of this teleological account of agency argued for in the work of Christine Korsgaard:

> A living thing is a substance so arranged as to secure the continuing existence of its own form. It does this in two ways: through nutrition, which enables it to preserve a continuing spacio-temporal stream of matter in its own arrangement or form, and through reproduction, which enables it to impose its form on other bits of matter. In other words, a living thing has a form that maintains matter in that very form. And that is its function. [...] Each kind of organism has its own specific ways of carrying out its nutritive and reproductive activities, or its own form of life. And we can identify it simply as the substance or entity that leads that form of life, or whose matter is organized in such a way that it maintains its form by living that form of life. Thus a dandelion is an entity that maintains its form through dandelion activities, such as spreading dandelion seeds on the wind, and a porcupine is an entity that maintains its form through porcupine activities such as defending itself with quills. In each case the function of the entity is simply to be what it is, to lead the kind of life it characteristically lives. (Korsgaard 2020:16-7)

The upshot — or so I argue — is that all self-reflexively organized teleological systems (i.e. all living organisms; systems that are organized toward the end of self-maintenance) are agents who perform actions which can be explained teleologically rather than purely causally. When a plant turns toward the sun in order to better

nourish itself, this is an action which can be explained teleologically rather than causally. In this way, not only are plants agents, but they perform actions which exhibit a peculiar independence from the causal chain. Specifically, the plant's actions can be explained perfectly well teleologically and without the need for appeal to causal history. While one way of explaining a plant's movements is causal, but there is an equal alternative explanation that's teleological.[23]

All living organisms — plants, animals, and humans — can have their actions explained by an appeal to purposes or reasons rather than an appeal to the determinist event-causal chain. This has the odd consequence of entailing that not only are plants agents, but agents who are originating first causes. This has interesting consequences for libertarian accounts of freewill which treat this as the singular important trait for explaining the special free nature of human agency. In the last section of the paper, I argue that plants and non-human animals have a form of agency that is nevertheless "lesser" and possesses a shallower sort of self-determination than human free agency. It is not that human actions are explained by justifying reasons instead of causes and explainable separately from the determinist event-causal chain that sets human agency apart from plant and animal agency like the libertarian claims (since that's true of plants and the lower animals too), instead it is the fact that our actions are fully self-determined in a self-aware way that plant and animal agency isn't. Plant and non-human animal agents do self-originate their movements in a way that's explainable separate from the event-causal order just like human agency, however plant and non-

[23] In fact, we might even imagine a world that cannot be explained or understood causally but only in this teleological way. It just so happens – perhaps – that our world is vulnerable to both forms of conceptualization and explanation (at least when it comes to living things).

human animal agents do not self-determine their own form of agency. Human agents, by contrast, originate not only their actions but their individual personal identity as agents in a way that makes them fully self-determining that plant and non-human animal agents are not. An individual human agent self-determines the particular self-reflexive form of life he or she lives whereas a plant or non-human animal only lives the general self-reflexive form of life of its species. To put it another way, each blade of grasses life is more or less the same and the teleological explanation of a plant's actions is determined by the generic blade of grass way of life shared by all blades of grass whereas each human's life can only be understood teleologically in terms of the particular life they've chosen for themselves (e.g. to be a good father, a loyal friend). Both are explainable teleologically in a way that is distinct from causal explanation however only human life involves the sort of self-aware way of life where the teleological explanation of a human action depends upon facts about the sort of life the human agent has chosen for him or herself. An individual human agent self-determines the particular self-reflexive form of life he or she lives whereas a plant or non-human animal only lives the general self-reflexive form of life of its species.

Which is to say, this paper is about the free agency typical of human beings as a special subclass of agency possessed by all life. This freedom isn't special because of its disconnect from the causal order and vulnerability to teleological explanation instead (or in addition) to causal explanation, since all living systems are subject to teleological explanations of their actions. What's unique about human agency is the way in which human agents are additionally in control of their own teleological organization.

To begin I turn to the problem of the disappearing agent for causalist theories of agency, how the agent-causal libertarian attempts to solve it, and then how a non-causalist theory of agency provides a better solution. I will then turn to metaphysical libertarian accounts of free will and how they end up missing what's unique about human agency which makes us think of ourselves as especially free or self-determining when compared to plants and non-human animals.

§2. The Problem of the Disappearing Agent

Causalist views of action are subject to a number of objections, one of which is the "problem of the disappearing agent". Velleman summarizes the problem in the following way:

> The flaw [...] is that the [causalist] story fails to include an agent — or, more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but nobody — that is, no person — does anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part. (Velleman 2000:123)

Harry Frankfurt (1971) attempts to solve the problem of the disappearing agent while still maintaining a causalist view by specifying certain mental states or events which can be said to "stand in" for the agent or serve as the agent's "representative". This is to say, if we ask the question: "How does a person figure into the causal order?" A Frankfurt-style answer isn't to treat the agent herself as an element in the causal order (as agent-causalists do) but rather to identify a kind of mental state which can be said to speak on behalf of the agent as a whole.

Frankfurt has us consider instances in which an agent's behavior has been caused by a mental state like an intention or desire but in which the agent

nevertheless seems to be "a helpless bystander to the forces that move him" (Frankfurt 1971:21). Frankfurt uses addicts as his primary example. Frankfurt then asks what's different about cases like the addict from cases where the agent does seem to be in control of his movements? Frankfurt comes to see the issue to lie with the way an agent can be "alienated" from the mental states that move him. The agent is motivated in these cases by a desire or intention which she does not herself "identify" with or "endorse". Frankfurt's suggestion is that the "agent's identifying with the motive that actuates him [...] consists in his having a second-order desire to be actuated by that motive" (Velleman 2000:133).

There is a worry, however, that an agent can also be alienated from his second-order desires. And this worry is not unique to Frankfurt's specific suggestion that identification consists in a second-order endorsement. Whatever we might take the mental act or state of identification to be, the mental act or state of identification will just be something that happens to an agent unless it is conceived of as something the agent does rather than a mental state or event within the agent causing their behavior. Bratman has memorably raised this objection to Frankfurt-style accounts by asking what makes the mental act of identification or endorsement anything more than "just one more wiggle in the psychic stew" that causally produces an action? (Bratman 2007:24). Which is to say, the Frankfurtian strategy of getting the agent involved in the causation of an action by specifying some special mental act or state of "identification" will merely push back the problem a level. Instead of asking how an agent participates in an action which is caused by one of her intentions, we are

now left asking how an agent can be said to participate in causing the mental act or state of her identifying with the motive which in turn caused this action.

Which brings us to a proposal from Christine Korsgaard. As with other critics of the causalist view of agency, Korsgaard takes the problem of the disappearing agent to deserve a serious answer. In her view, it is a fundamental characteristic of actions that someone does them: an "action is not just something that happens in or to the person's body or his mind but rather is something that he as a person does." (Korsgaard 2008a:160). What, then, would it take for the agent herself — i.e. the whole agent rather than merely a mental state possessed by the agent — to cause an action? Korsgaard makes the following proposal:

> What makes an action mine, in the special way that an action is mine, rather than something that just happens in me? That it issues from my constitution, rather than from some force at work within me [...] [or a] a law imposed upon me from without. (SC §8.1.3)

And,

> To see a movement as an action, [...] we must see it as a movement attributable to the [agent]'s form. Since the [agent]'s form is what unifies her into an individual object, her form is not merely something within the [agent]. So when the [agent]'s movement can be attributed to her form, it is the [agent] herself, the [agent] as a whole, who moves. (SC §5.5.4)

The agent as a whole is the cause of a movement or effect just when the movement or effect is attributable to the agent's identity or form. The notion of "form" here is Aristotelian and amounts to the agent's identity or persistence conditions — i.e. her essential properties. A statue and the lump of bronze that composes the statue possess different individual persistence conditions (specified by their criteria of diachronic identity). The lump of bronze which composes the statue existed before the statue itself and can survive changes which the statue cannot. Therefore, the statue and the

lump of bronze that composes the statue are different things even while they happen to spatiotemporally coincide for a time. The conditions which determine when the lump of bronze came to be shaped such as to compose a statue that didn't exist before and what changes the statue can undergo without ceasing to exist are the statue's *form*.

But what does it mean for an agent's movements or effects to be "caused" or "attributable" to her "form"? Presumably, it isn't correct to model the relationship between an agent's form and her effects or actions on the model of one billiard ball striking another. An agent's form is not a physical cause like a billiard ball that can relate to an effect or action in the manner a striking billiard ball can. So, then, what could it mean to say a movement or effect is attributable to an agent's form and so attributable and caused by the agent herself? The model of causation we must look to here is that of substance causation rather than event causation. And this is precisely the matter I turn to in the following section.

§3. Agent-causal Libertarianism

In metaphysics, there has long existed a debate regarding the nature of causal relata — are they events or substances? Consider the following two ways of describing a causal event:

The explosion of the bomb caused the collapse of the bridge.

The bomb caused the collapse of the bridge.

The former is an example of event causation—the causation of one event by another event—and the latter exemplifies substance causation—the causation of an event by an individual substance. A 'substance' in this sense is a persisting object or substance possessing various properties and persisting through change. There is debate whether one or the other form of causation is reducible to the other (i.e. whether substance causation is analyzable in terms of event causation or vice versa). A non-reductionist view holds that both species of causation exist without either being reducible to the other. However, the standard view in the literature is reductionist; specifically, that substance-causation reduces to event-causation.

A particular form of libertarians known as "agent-causalist libertarians" are centrally concerned with arguing that — *at least in the case of agents* — substance causation is an importantly distinct form of causation that cannot be reduced to event causation (i.e. unlike statements about bombs blowing up bridges). Why? For one, if substance causation were true — *at least in the case of agents* — then we can side-step the problem of the disappearing agent. Actions are caused by the agent's themselves via substance causation rather than merely by some mental state or event occurring in the agent. But, more than this, the idea of substance causation purports to offer us a way out of the determinist's dilemma.

Insofar as we view an agent's actions as caused by a decision (or choice, intention, or other mental state of an agent) understood in event-causal terms, we face a dilemma. Either the agent's decision was itself caused by prior events which trace backwards to events before the agent was born in a way that undermines the notion of the agent as the originator of her action (rather than just a conduit in a long running

event causal chain that happens to run through her) or, alternatively, the agent's decision lacks any prior event cause. Traditional libertarians grab the second horn of this dilemma by asserting that an agent's decision to act lacks any prior event cause, but it's long been recognized that this doesn't secure traditional libertarians the result they aspire to. If an agent's decisions or actions are uncaused, then their actions do not seem to originate from the agent at all and furthermore would seem to be just random. Merely arguing that the actions of agents are not causally determined by previous events doesn't in the end amount to what we mean by free action. If an action is just the result of random indeterminacy, that doesn't constitute free agency. This is a problem traditional libertarians face and regarding which agent causalist libertarians think they can do better. For the agent-causalist, these gaps in the event causalist order are filled by the agent herself by means of substance causation. The agent, as a substance, isn't itself an event and so does not stand in need of a previous event cause (or so the agent-causalist libertarians claim):

> Since a substance is not the kind of thing that can itself be an effect (though various events involving it can be), on these accounts an agent is in a strict and literal sense an originator of her free decisions, an uncaused cause of them. (Clarke 2000)

By beginning a causal chain with a substance instead of an event agent-causal libertarians claim they have passed through the horns of the determinist's dilemma. They claim to be offering us an explanation of our actions by appeal to something outside the event-causal order — in particular, *us* as substances — and so something which does not thereby end up being explained by the event-causal chain back to the beginning of time, and yet something which can (supposedly) still cause events and

— moreover — do so not merely randomly but on the basis of the agent's justifying reasons.

The perennial criticism faced by agent-causal libertarians has always been that the view is obscure at best and incoherent at worst. Which is to say, the most repeated criticism of advocates of agent-causal libertarianism has never been this-or-that counterexample or objection but instead just that no one has been able to put forth the view in a way that can be made sense of. Even one of the view's most prominent defenders — Richard Taylor — is reduced to concluding one of his works with the following defeatist thought:

> One can hardly affirm [an agent-causalist] theory of agency with complete comfort [...or…] wholly without embarrassment, for the conception of agents and their powers which is involved in it is strange indeed, if not positively mysterious. [...] Perhaps here, as elsewhere in metaphysics, we should be content with discovering difficulties, with seeing what is and what is not consistent with such convictions as we happen to have, and then drawing such satisfaction as we can from the realization that, no matter where we begin, the world is mysterious and that we who try to understand it are even more so. This realization can, with some justification, make one feel wise, even in the full realization of his ignorance. (Taylor 1963:53)

The idea Taylor is expressing here appears to be that the picture which agent-causalist libertarianism provides only enables us to have some vague, slightly incoherent grasp of the place of free will in the world that is still worth having even if ultimately we can't fully make sense of it. In what follows, I aim to do better. It is perhaps the case that the view I defend isn't one that the traditional agent-causalist would accept, but I do think it makes sense of many of the intuitions which drive philosophers toward agent-causalism. To that end, I begin with a list of a number of objections which one could raise toward agent-causalist libertarianism and how the view of agency I

developed in the other work can successfully answer these objections in ways which traditional agent-causalists leave mysterious.

Aside from the charges of incoherence and impenetrability, there are (at least) three outstanding problems with agent-causal libertarianism. By addressing each of these three problems, I hope to illustrate how my proposed account of agency better fulfills the promise of traditional agent-causalist libertarianism. First, traditional agent causalism requires breaks in the event-causal order regarding which agent or substance causation fills in the gaps. For some agent-causal libertarians this involves an implausible reliance on quantum indeterminacy in the event-causal chain leading to our actions to open up the window for our agent-causation to be exercised. Second, a substance cause would itself seem to be the product of events which bring it to exist. So even if an action is the result of a substance, it seems to still belong to the event-causal chain given that the existence of the substance itself is a product of the event-causal order. Lastly, agent-causalist libertarians often argue that their scheme allows actions to be explained by the agent's reasons rather than prior event-causes. However, why would attributing an action causally to the agent herself as a form of substance causation enable us to view the action as explained by the agent's reasons? It's at least not immediately obvious why attributing the cause of actions to the agent as a substance entails that these actions are now open to rational explanation instead.

Agent-causalist libertarians offer different answers of varying quality to the questions I've posed above, but I won't survey each of them here. As I've said, I instead want to explain why the account of agency developed in the other work offers us a perspicuous way of making sense of the intuition behind agent-causalist

libertarianism in a way that sidesteps each of these challenges as well as providing us with a more penetrable account of how agents themselves can be the causes of their actions.[24]

§4. Answering (some of) the Problems with Agent-causal Libertarianism

At least certain parts of the world are open to conceptualization as teleologically self-reflexive systems. This form of conceptualization may not be possible for rocks or artifacts like mouse traps or cars, but it is possible for living organisms like plants, horses, and human beings. Specifically, the behaviors and parts of a living organism like a horse can be conceptualized as purposeful and directed toward the preservation, self-maintenance, and promotion of, for example, the horse and the horse's life. Moreover, the horse itself can be conceptualized as having the function or purpose of operating to preserve, maintain, and promote the component parts that make it up. That's the sense in which the system is "self-reflexive" and which differentiates living systems from other sorts of teleologically organized systems like alarm clocks or cars. Alarm clocks are teleologically organized toward a purpose, but the purpose is imposed on it from the outside. A well-functioning alarm clock can only be good or useful for us, it is not good for the alarm clock itself for it to function well. By contrast, because a horse is self-reflexively teleologically

[24] Another problem is this: 'A substance is a persisting entity and so how can a decision at some specific moment in time be attributed to the substance which existed long before the decision it is causally responsible for?' This objection is I think the easiest to handle of those I've listed and its answer isn't particularly dependent upon Korsgaard's account of agency — which is why I've relegated this further objection to a footnote. Here is my answer: A substance has as its form a nature to, for example, 'do x in y conditions'. While the presence of y conditions is part of the explanation for what happens and what an agent does, it isn't the complete explanation. Part of the explanation is the agent's constitutive form and so an appeal to the agent. Not all agent-libertarians are happy with this response and there's a longer discussion that has been ongoing regarding it, but since engaging with that literature doesn't help illustrate the view of agency I'm advancing I won't say more about it here.

organized, a well-functioning horse might be useful for us (because we can ride it or sell it) but it's also good for the horse itself. A well-functioning horse is one that contributes to the continuation of its horse behaviors and horse component parts functioning not for some external purpose but instead self-reflexively to continue to keep the horse and its horse behaviors and horse parts in existence and well-functioning. To put it another way, the activity of a machine only makes sense in terms of the purpose given to it by an outside user whereas self-reflexively organized teleological systems — i.e. living organisms — can be understood as functioning for purposes wholly of their own.[25]

When we conceptualize the world in terms of self-reflexive teleological organized systems, this also opens up a new separate form of explanation for phenomena distinct from event-causal explanations. Instead of explaining why an antelope stots in event-causal terms by appeal to the neural-firings in its brain, I can cite the purpose of the antelope's behavior. The antelope is stotting to scare away the predator it detects and, ultimately, to thereby promote its own continued existence as a thing that self-preserves by — among other things — stottling to scare away predators. Which is to say, in response to the question "Why did the antelope stottle just then?" there is both an event-causal explanation appealing to brain-states in the antelope as well as a *sui generis* teleological explanation which gives the reason for why the antelope stotted by citing the purpose such a behavior has in the antelope's

[25] It is also possible that ant colonies, computer viruses, stars, and ecosystems generally can be conceptualized in this manner. If that is the case, then such things are alive and agents as well. However, I need not take a strong stance on these cases here. I do speak about ant colonies at length in this dissertation and I do think they are both living organisms and agents, but I am less sure about, e.g., computer viruses.

way of life. These are not competing explanations that crowd one another out such that only one can be true, rather for self-reflexive systems there is simply an independent way of explaining what happens with regard to them in a way that doesn't cite event causes but instead cites purposes or reasons.

This picture offers us both a way to understand the manner in which agents can be the causes (or explanations) of occurrences rather than the event-causal order as well as why reasons play a role in agent causation. The explanation for why the antelope stotts is by appeal to the nature or form of the antelope. An antelope has the form of something whose behaviors and parts are teleologically organized to preserve itself as an antelope — that is, something engaging in behaviors and possessing parts that are teleologically organized to preserve itself in the manner distinctive to an antelope. And so this alternative teleological form of explanation that sits alongside and separate from explanations which cite the event-causal history of an event is one that cites the agent herself and by appeal to her form — what makes her the sort of agent she is. Furthermore, it is an explanation which cites the agent's reasons. I explain the antelope's stotting behavior here by giving the reason for it in the way of life that defines what it takes to be an antelope.

Traditionally, agent-causal libertarianism takes an agent-causal explanation of an agent's movement and an event-causal explanation of an agent's movement as mutually exclusive. If there is an event-causal explanation of an agent's movement by appeal to, for example, the neural-firings in the agent's brain then that movement wasn't caused by the agent but instead by the whole event-causal chain of causes leading back to the beginning of time. To make room, then, for libertarian freewill

and agent causation, traditional agent-causalism argues in favor of the view that there are "breaks" or "gaps" in the determinist event-causal order which agent causation fills in the gaps regarding. In the breaks in the determinist causal order resulting from — for example — quantum mechanics, there is room for an agent's movements to be caused by the agent herself rather than prior events. Somehow this causation by the agent isn't random, but on the basis of reason (but how this could be so seems rather mysterious). Furthermore, what it could mean to say the agent is the cause of a movement herself as a substance is a notion that's difficult to render clear. However, the approach to agency I developed in the previous chapter offers us a form of agent-causal libertarianism which escapes these problems. On this proposal there is an event-causal view of the world which is fully determined, however once we switch to a teleological view of the world then the actions an agent performs are explained via the place of a movement in the teleologically self-reflexive order of the agent's form of life. With self-reflexive teleologically organized systems we can take a certain particular sort of rational or intentional stance toward a thing and view them teleologically to explain their movements and components in a way that we can't with other things. This doesn't "crowd out" event-causal explanations for the movements an agent's body performs; event-causal explanations are simply a different sort of explanation than the teleological stance we are concerned with when we speak of agents and the explanation of their actions. There is an event-causal view of the world which is fully determined, however once we switch to a teleological view of the world then the actions an agent performs are explained via the place of the movement in the teleologically self-reflexive order of the agent's form of life.

This approach to agent-causal libertarianism has its roots in "dual-aspect" or "two-standpoint" interpretations of Kant. According to Kantians who embrace the dual-aspect view (most prominently represented by Henry Allison's 1983 *Kant's Transcendental Idealism*), when confronting the world there are two perspectives which offer us two-aspects of one and the same world which cannot be cognized separate from one of these two equally legitimate standpoints. Here is how Korsgaard summarizes the idea which she also endorses:

> I take the view I am advancing to be a Kantian one. Kant believed that as rational beings we may view ourselves from two different standpoints. We may regard ourselves as objects of theoretical understanding, natural phenomena whose behavior may be causally explained and predicted like any other. Or we may regard ourselves as agents, as the thinkers of our thoughts and the originators of our actions. These two standpoints cannot be completely assimilated to each other, and the way we view ourselves when we occupy one can appear incongruous with the way we view ourselves when we occupy the other. As objects of theoretical study, we see ourselves as wholly determined by natural forces, the mere undergoers of our experiences. Yet as agents, we view ourselves as free and responsible, as the authors of our actions and the leaders of our lives. The incongruity need not become a contradiction, so long as we keep in mind that the two views of ourselves spring from two different relations in which we stand to our actions. When we look at our actions from the theoretical standpoint our concern is with their explanation and prediction. When we view them from the practical standpoint our concern is with their justification and choice. These two relations to our actions are equally legitimate, inescapable, and governed by reason, but they are separate. Kant does not assert that it is a matter of theoretical fact that we are agents, that we are free, and that we are responsible. Rather, we must view ourselves in these ways when we occupy the standpoint of practical reason — that is, when we are deciding what to do. (CKE 377-8)

Kant did not have the Aristotelian-teleological flavor to what he took to be involved in occupying the practical standpoint, but the idea of this teleological conceptualization of agents as something that follows from a certain different standpoint we can take on the world — and moreover one that is necessary for action (more on that later in this paper) — is a Kantian idea under one interpretation of

Kant. Agent-causation doesn't require a break in the causal order as traditional agent-causalist libertarianism claims, but instead a different standpoint we have available to us when it comes to living organisms.

Which brings me back to reconsider the first problem I raised with agent-causal libertarianism. The first problems I raised for the traditional agent-causal view was that it implied an implausible break in the determinist causal order for the agent to exploit and exercise its libertarian agent-causal power. As is clear from what I've already said, that is not part of the picture of agency I've developed in the previous chapter. Instead, there are two equally legitimate standpoints with which we might consider systems which are susceptible to being conceptualized as self-reflexively teleologically organized. Under one way of viewing the world, it is understood in event-causal terms. There is no break in this causal order for human free decision-making, it's fully deterministic and event-causalist. Under the other way of viewing at least certain parts of the world, there are teleologically organized self-reflexive systems which can be understood in a totally different way from event-causalism. *It isn't that there are event-causal gaps in the determinist order which agent-causalism fills in, but rather that viewing oneself and the world as filled with agents is just a wholly different way of conceptualizing the world.*

But the second worry I've raised regarding traditional agent-causalism is that the existence of substance causes are themselves the result of the event-causal chain. In some sense, the explanation for why an animal exists can be explained simply by giving how the animal's behaviors self-maintain and continuously re-create itself. A living organism is a system teleologically organized to keep itself in existence by

self-preservation and so by explaining how a living organism works to continue its existence can explain why it exists at least in part — *and perhaps even maybe fully*. I say "and perhaps even maybe fully" because suppose the world we lived in was as Aristotle hypothesized. The world had no beginning and has existed stretching eternally backwards. Individual living creatures have either always existed re-creating themselves through self-maintenance (perhaps God or the universe itself is like this). However, it does not seem like we live in the eternally existing world that Aristotle took us to occupy. The existence of self-reflexive systems has an event-causal history which explains how such things came to exist — even if once they do exist teleological explanations can be provided for what they do and how they continue to exist. Which is to say, we can't do as Aristotle did and explain their existence solely by appealing to their self-maintenance and reproduction since things like birds or grass or human beings haven't just existed forever back into eternity but evolved and came into existence. So does this mean that ultimately teleological explanations trace back to a beginning which reverts back to needing to be explained by the event causal order? In one sense "yes", in another sense "no".

Here I find an analogy with explaining what occurs in a film helpful. When characters in a film do something, we can often explain their actions within the internal narrative of the film. However, sometimes characters act or events transpire which we cannot make sense of from within the narrative of the film. We have to explain what happens as due to things like "well the writers needed an excuse to bring back into the story a character played by a famous actor who will help the film sell tickets" or "the characters didn't solve their problem in the obvious way available to

them because then the story would end too abruptly". These are external rather than internal explanations for what happens in the film, and usually when we are unable to provide internal explanations for what occurs in the film and are forced to appeal to external ones like these we consider that bad writing. There is a sense, however, when we try to take a film and characters and a narrative seriously that we assume or do our best to try and understand or view the film as fully comprehensible through internal explanations. We put out of mind the odd sudden appearance of a character who has no motivation for suddenly showing up because doing so breaks our ability to engage with the film where we understand events to be happening due to internal narrative reasons rather than external movie-making pragmatic reasons. It may be true that we can see that the character's sudden unmotivated re-appearance in the story is due to the external explanation that the character is played by a famous actor who the writers needed to show back up in order to sell tickets for the movie, but that isn't an explanation we can give from the perspective of someone engaging the story from the internal perspective.

Here's the analogy: while the film itself exists for reasons external to the film's narrative, and sometimes the events within the film cannot be explained internal to the film's narrative, this does not obviate the importance or usefulness of explaining things internally to the film. If a character played by a famous actor shows back up into the plot, perhaps there is both an excellent internal to the narrative reason for this to happen even if we are aware that there is a double-explanation outside the narrative.

Self-reflexively teleologically organized systems have behaviors which are explainable from within the narrative of the self-maintaining life of the organism separately from the event-causal history. For systems of this type, there is an alternative form of explanation which is separate and sui generis from the event causal history and which instead appeals to the agent's form (i.e. substance causation). This explanation within the narrative of the self-maintaining life of the organism explains both the actions of the parts and the existence of the whole teleologically rather than by citing the event-causal history.

The last worry I've raised regarding agent-causal libertarianism is the idea that reasons explain an agent's actions even when event-causes do not. Teleological explanations of actions come to be grounded both in the agent's form as well as the reasons for the agent's action. When a hungry lioness hunts, it does so to find food. Her movements can be rationalized according to the place of this movement in the form of life of the animal. By engaging in this behavior the lioness both instantiates herself as a lion as well as performs a movement which is rationally explicable. The movement is both thereby subject to the dual explanation in terms of reasons as well as being attributed to the form of the agent by being located as an action in the agent's teleological form — i.e. part of what constitutes the agent as a lioness agent. The action is explainable by reason because — qua action — it is a movement teleologically explainable in the life of the lionesses way of life. It is an action attributable to the lioness because it is by performing such actions that the lioness qualifies as living the life of a lion and constitutes herself as a lion. Just as why you might be able to explain why a character returns into the story both by appeal to the

internal story reasons for why it happens as well as the external pragmatic reasons of the writers making sure to give a famous actor screen time in order to sell ticket, both explanations are warranted but distinct and sit alongside one another offering us two different ways to explain and understand what happens.

Actions are unique in that they are justified by reasons. For a causalist about action (i.e. someone who sees reasons as causes) this involves a mental state cause of an action also being one that rationalizes the action. But for an anti-causalist the rationalization of an action does not work this way. Anscombe suggests that instead we justify actions by appealing to the broader actions of which a particular movement is a sub-action; a rational justification does not to cite a prior cause (not even a mental state like an intention which might seem to be what rationally justifies the action), rather a rational justification cites the broader unfolding action which one is engaged in. To illustrate:

> We often say of an individual who acts at a particular moment: "By doing A, she is φ-ing" (where φ-ing is some temporally extended activity that unfolds over time). [...] Such is the relation of egg-breaking and egg-mixing to omelet-making. [...] So when an individual cracks an egg on an occasion in which she is omelet-making, her action is not distinct from the omelet-making itself. "Making an omelet" is every bit as true a description of what she is doing as "cracking an egg." However, when the individual adds some vegetables to the mix a few moments later, she is also making the omelet--the very same omelet. (Schofield 2012:95-6)

Not that if this agent were asked why they were breaking these eggs or adding some vegetables to the mix, they *could* give as their reason "I'm making an omelet." This isn't citing some prior mental state which is causing both of these actions, but rather simply a description of what the agent is doing on both occasions. On both occasions — i.e. breaking these eggs and adding vegetables — the agent is making an omelet.

And when asked for his reason, the agent simply describes what he's doing (albeit in broader or fuller terms). The agent's reason is not a prior intention which caused them to act, it is just a description of what they are in fact now doing.

But while an agent *could* give as his reason "I'm making an omelet", they could also just as well answer "I decided to have an omelet for breakfast" or cite his hunger as his reason or say "I'm hungry and I just like omelets." The last of these responses isn't really different from what I've already discussed. When an agent responds "I'm hungry and I just like omelets" he's in part answering "I'm making an omelet" but more than that he's describing how in doing so he's engaged in an ever broader action which making an omelet is only a sub-part. Namely, he's making an omelet as a sub-action in the broader activity of satiating his hunger and appeasing his taste preferences. In general, if someone were to say "I'm hungry and I just like omelets" I'd take them to have a general project of taking seriously their bodily desires and seeking to satisfy them (both of hunger and taste).

The first response — i.e. "I decided to have an omelet for breakfast" — is more complicated. That I take to be little more than equivalent to answering ""I'm making an omelet." This is because citing a decision or intention is not — on Anscombe's view — a causally efficient mental state. An intention is rather just a description of what one is doing. Let me explain.

But if an intention isn't a prior mental state which causes actions (and imbues them with their intentionality), what are intentions? Anscombe argues that an intention as a "location on a spectrum of unfolding action that fits an event to be described using concepts of intention in the first place":

> Picturing a line, on the far left will be found pure intentions, defined as cases in which the agent intends to do something but hasn't yet done anything else in order to do that. Moving rightward, the agent has more worldly deeds to show for his intentions: if he is described as φ-ing or as intending to φ, then, at this point, it will be correct to say that he is also doing various things in order to φ, or because he intends (wants, aims) to φ. At the far right, his performance is fully unfolded and finds description in the past tense: 'He φ-ed' (or 'He has φ-ed') will now be true, and not merely—what holds anywhere between these end points—'He was φ-ing.' (Moran and Stone 49:2011)

Which is to say, the following expression — "X is intending to φ" — is really an expression of the fact that the agent is performing an early sub-action in the carrying out of φ rather than an occasion wherein X has some particular sort of mental state that causes the action (as claimed by the causalist view of agency).[26] [27]

So an action is a movement justifiable by appeal to a larger unfolding action of which it is a part. The worry we face now is one of infinite regress. If every action is explained in terms of it being a sub-action in a larger-action, where does this chain of explanations end?[28] Korsgaard's view (as discussed in the previous chapter) is

---

[26] Let's return to the problem which prompted me to address Korsgaard's allegiance to the Anscombian view of action in the first place: deviant causal chains. An advantage of viewing actions in this Anscombian way is how it side-steps the problem of deviant causal chains.# Once we adopt an Anscombian view, we interpret actions holistically. And a movement caused in the way described by Davidson's deviant causal chain example would not holistically be interpreted as the same sort of action as it would be interpreted to be in the normal case.

[27] Now, one peculiarity of Anscombe's view is that it implies that it is never true that "X is intending to φ" unless X is engaged in carrying out some sub-action toward completing φ. But isn't it the case that I can decide to go for a run tomorrow and — at the moment I decide and form this intention and for some time afterwards — do nothing toward carrying out that action since the time had not arrived? Couldn't it be true that I decide to go for a run tomorrow and there is nothing available to me at present to currently start carrying out that decision? Moreover, suppose I am struck by a bus and killed before I begin to take any steps toward carrying out my decision; does this mean I never intended to go for a run tomorrow even after deliberating carefully and conscientiously for a long-time before settling on that course of action and just not yet having any way to act to progress it?

[28] Elijah Milgram poses this problem in a slightly different manner. Milgram asks how it is possible for every action we perform to have an infinite series of smaller subactions which belong to it. Milgram writes: "[N]otice that this metaphysics has consequences that are strictly incredible. We are faced with an infinite regress, downward, in the composition of intentional actions, entailing that any action is composed of infinitesimally tiny actions. [...] But human beings are small-finite creatures, and so there could not be that much attention and control to go around." (Milgram 2009:173).

uniquely situated to handle this problem. The solution to the regress of nesting actions in larger actions is the idea of self-reflexive activity. The chain of explanations need not end, it can circle back on itself. The idea is that we escape the regress through circularity (hopefully of a virtuous rather than vicious kind). If an agent's form is a self-reflexive activity, then every action an agent performs can be nested within the broader activity of its specific way of living (i.e. the specific form of self-maintenance it embodies). That is, an answer to this regress worry can be found in Korsgaard's claims that certain types of things (ant colonies, giraffes, human beings) have a self-reflexive, self-maintaining activity as their form.

Again, the agent-causal libertarian view is often accused of being obscure and somewhat mystifying about how it is supposed to work. I take Korsgaard's anti-causalist picture of agency to offer us a less obscure manner to understand the idea. Teleological explanations of actions come to be grounded both in the agent's form as well as the reasons for the agent's action. When a hungry lioness hunts, it does so to find food. Her movements can be rationalized according to the place of this movement in the form of life of the animal. By engaging in this behavior the lioness both instantiates herself as a lion as well as performs a movement which is rationally explicable. The movement is both thereby subject to the dual explanation in terms of reasons as well as being attributed to the form of the agent by being located as an action in the agent's teleological form — i.e. part of what constitutes the agent as a lioness agent.[29]

[29] There is the further question of how an action can be attributed to an agent's persistent form and yet be caused by the immediate events preceding it. However, this seems answerable by the idea that there are triggering events which lead to actions via the agent's persistent form — as in, when the agent in such-and-such circumstances they will hunt to find food. A lioness hunts for food when hungry is part

The upshot is that all living things — all agents — produce effects through a non-reductionary form of substance causation and so are not determined by the prior event-causal order.[30] When plants or lionesses or humans act, it is attributable and explainable by both appeal to the reasons the agent acts and by appeal to the teleologically ordered form of the agent. All life has self-reflexive teleological forms and so are agents — at least in the broadest sense of "agent". In a sense, all living organisms are free agents since their actions possess intentionality and their actions are subject to rational explanations grounded in their form rather than the event-causal chain.[31] Nevertheless there is a deeper sense of free action which only human action possesses.[32] It is to that topic I will come to address later in this paper. But let me begin to introduce the notion of human self-aware, free agency by beginning in

---

of her form but it has event preconditions even if her doing this action is rationally explained rather than part of the event-causal order. What the lioness does in response to these circumstances is explained by her form and the reason-explanation stemming from her form, it is not a causal result of the events leading up to her action.

[30] I am tempted to add here "at least when conceptualized teleologically" but that might not be accurate. There aren't really agents at all except when we view the world through the teleological lens where such agents are not determined by the previous causal order. When we view the world through the causal lens there is instead just another mechanistic system that is to be explained in event-causal terms.

[31] This self-preserving form of operating is the result of causal history, evolution, and selected-effects functions, however the conceptual framework which we apply to a living organism is as it is now. Selected-effects functions and causal-evolutionary history might be relevant to understanding how a self-reflexive organism came to exist when considering the world from a causal conceptual standpoint, however once such a self-reflexive organism has come into existence now there is an alternative teleological conceptualization possible where its origin can be explained solely within that teleological conceptualization of the world. It doesn't feel like a complete explanation precisely because we can see that by switching to the causal-historical standpoint and explain the world in that alternative way also.

[32] Korsgaard notes a similar distinction between these two senses of free agency: "[T]here are actually two senses of autonomy or self-determination. In one sense, to be autonomous or self-determined is to be governed by the principles of your own causality, principles that are definitive of your will. In another, deeper, sense to be autonomous or self-determined is to choose the principles that are definitive of your will. [...] Every agent, even an animal agent, is autonomous and self-determined in the first sense, or it would make no sense to attribute her movements to her. Only [...] human agents[ ] are autonomous in the second and deeper sense [...] [and thereby also] agents in a further, deeper sense. For we do not merely determine ourselves in the sense that we act from the principles of our own causality—we determine ourselves in the deeper sense that we choose the principles of our own causality. It is up to us to decide what we will do for the sake of what." (SC §6.1.1).

the next section with summarizing the idea of plant and non-human animal agency generally. From there I will turn toward the specific form of "deeper", free agency possessed by human beings.

§5. A Process Ontology

One might reasonably wonder, what does it mean for something to have a "way of life" or "activity" as its form? To steal a wonderful example from Paul Schofield (2012), an ant colony is similar to an agent in that an ant colony has an activity as its form. Particular ants in the colony move dirt in such-and-such way in order to maintain a tunnel. Other ants scout for food in order that other worker ants can bring that food back to the colony for storage. Still other ants occupy the role in the colony of feeding and caring for the queen who herself occupies the role of reproducing further members of the colony. Which is to say, an ant colony is no mere mereological sum of individual ants which just happen to be spatially near one another. An ant colony is a distinct object which is composed of these individual ants functioning together in such-and-such way just as a bronze statue is a distinct object which is composed of a lump of bronze shaped in such-and-such way. An ant colony is not some metaphysically queer object, but it is something over and above the mereological sum of the ants that make it up.

Groups of organisms have a unifying form and — moreover — their form is an essential or vital activity constitutive of the group's identity as a numerically distinct existing thing rather than a mere mereological collection of a number of

distinct individuals.[33] Some individual ants must have existed before a colony can exist, but the colony itself is not a phase which the ants enter into by engaging in certain activity together. The ant colony is something distinct from the individual ants that compose it and that comes into existence once the ants start to engage in a certain activity together. Moreover, the ant colony will immediately go out of existence once the activity grounding its existence ceases even if all the individual ants who constituted the ant colony may continue to live on after the colony disappears. An ant colony is an object which exists only insofar as a certain activity is being engaged in and has that activity as its form.

In a certain sense, ant colonies can be said to be "self-constituting". Individual ants are disposed by instinct to behave in a certain way, and by doing so they collectively come to constitute a group activity which the ant colony as a whole performs and supervenes upon. The ant colony itself is self-constituting in the sense that it only comes to have the "form" or "identity" or "character" of an ant colony and thereby exist via the ant colony behaving as an ant colony. The existence of the ant colony does not supervene upon the dispositions of the ants to interact collectively in

[33] Suppose, for example, we imagine a contingent of ants from a particular colony leaving to start a different colony some distance away. After enough time has past that now both colonies are composed of a few generations later of new ants, one of the members of the offshoot colony happens to comes across the original colony while scouting for food. This scout proceeds to enter the original colony unchallenged given the fact that the mechanisms which normally signal an intruding foreign ant fail in this case given the scout ant's family history to this colony. The scout ant then proceeds to retrieve food from the original ant colony's storage just as it he had discovered a cache of food on the surface. He then proceeds to return the food to his home colony. Now, while this ant is in the hive of the original colony, is he a member of that colony? He's genetically identical to members of that colony, but this doesn't seem to be enough. He is spatially near all the other members of the colony, such that if we view the collection of ants in this region of space as a mere mereological sum of ants rather than a colony then he would be a member of that mereological collection of ants. But this doesn't seem either to make him a member of that colony. To be a member of the colony, he must participate in functioning or activity which we take to be constitutive of the colony.

a certain way, it supervenes upon the actual collective behavior as it occurs. An ant colony comes to constitute itself as an ant colony by behaving in the manner constitutive of a living ant colony. Which is to say, an ant colony comes to exist by that very ant colony doing ant colony things.

Our notions of a 'baseball bat' and a 'pitcher' are similar to our notion of an 'ant colony' in this regard — i.e. in that the form of each is an activity. Or, to be more accurate, our notions of a 'baseball bat' and a 'pitcher' are polysemous such that in one of their senses the form of each is an activity. In one sense, a bat is only a bat when it's actively serving as one in a game of ball. In another sense, an object that isn't even ever used to play baseball can be a baseball bat. Call these the "active" and "static" senses of the concept of a baseball bat.

Similarly, we might speak of the team's pitcher as the player assigned on the roaster to play that role in games. But in a particular game, due to an injury, the coach may need to substitute in the first-baseman as pitcher. Hearing that the pitcher was injured and had to be replaced, I can ask you "Who's the pitcher?" and you can answer "The first-baseman." This makes sense only if we are using these concepts in different senses. I mean "pitcher" in the activity sense and you use the notion of "first-baseman" statically. Now, note that in the activity sense, being a pitcher is an activity. I must be engaging in the essential activity definitive of a pitcher in order to be one, as soon as I stop, I stop being a pitcher in this active sense. I make myself into a pitcher by engaging in the activities constitutive of a pitcher. Perhaps I am not a pitcher if we take to the field and only throw the ball once. That isn't a baseball game,

not even a short one. But over time our behavior comes to constitute a baseball game and I come to be a pitcher throwing pitches.

I am unsure what Korsgaard exactly would say here. I take it that an ant colony should and would qualify as a living being by her Aristotle-inspired proposed notion of "life". And an ant colony comes into existence by that very ant colony engaging in ant colony activities. It is a system which has as its overall purpose promoting its own components in maintaining itself existing as an operating ant colony. That is, an ant colony's purpose as a whole is to support, promote, and preserve the worker ants, the queen ant, the reproduction and replacement of its population, and so on which themselves have purposes with respect to how they enable the colony as a whole to perform this self-reflexive function. Like a pitcher or a baseball bat in the active sense, an ant colony is really something that has engaging in a certain activity as its form. An ant colony in some static sense is just the collection of ants that makes it up (just like a pitcher in a static sense is whoever is designated to play the pitcher role on the team and a baseball bat in the static sense is something made to play the role of a baseball bat in the active sense). But the static sense of our notion of an ant colony is derivative from the more fundamental idea of an ant colony in the active sense. An ant colony brings itself into existence by operating as an ant colony. A number of individual ants engage in behavior which collectively instantiates an ant colony's behavior and in virtue of this an ant colony

comes to exist over and above these individual ants in virtue of there being an ant colony operating as an ant colony.[34]

This makes sense of the Anscombian view of agency. When individual ants perform actions which are not merely individual but embodying the ant colony doing something, their actions can be justified in terms of higher actions. The highest action in this hierarchy is the ant colony's way of life and even this action can be in turn justified in how it is the ant colony's way of supporting itself, promoting, and performing all the component actions that make up its way of life.

§6. From Plant Agency to Human Agency

Living agents have a particular manner of self-maintaining activity as their form. A living agent consists of a number of activities teleologically organized toward keeping itself in existence. And what is it for a living agent to keep itself in existence? What is the living agent that these activities are directed at preserving? Precisely something which self-maintains in the particular manner distinctive of that particular form of life. This is the unique sort of self-reflexive teleological organization unique to life and particular sorts of agents.

Living beings are objects that have the function of maintaining their own forms. That is, that "[t]here is a kind of self-referential character to an organism's

[34] The baseball analogy is useful in capturing how the existence of individual parts and the existence of the whole are interdependent and come into existence together. However, a baseball game is disanalogous from an ant colony and other forms of life in that a baseball game isn't self-reflexive and the thus the game isn't responsible for creating and self-maintaining itself when viewed from the teleological-intentional conceptual standpoint. Which is to say, the baseball game analogy is only meant to capture one aspect of the metaphysics of living organisms (i.e. the way the parts and whole stand in a holistic existential relationship) but isn't wholly analogous which is why a baseball game isn't a living organism on my account (but an ant colony is).

functioning, for its function is more or less to preserve a certain way of functioning, the way that is characteristic of its kind, and nothing more." (Korsgaard 2018:20). This function that living creatures possess is not due to historical, origin-based functions (i.e. SE-functions), but the result of ahistorical system functions (i.e. Cummins' CR-function) when they are applied to a self-reflexive system.

However, as Korsgaard takes herself to follow Aristotle, I will follow Korsgaard in distinguishing several key sorts that "life" divides up between. Plants self-maintain in many different ways, but there is one key manner of self-maintenance no plant engages in. Specifically, they do not do so in the unique way animals do. Animals also self-maintain in many different ways from one another, but what they all share in common is that animals pursue self-maintenance via representation of the world around them, which plants lack:[35]

> [W]e can also draw broad distinctions among types of life forms. Plants are the basic form of living organism, characterized simply by the powers of nutrition and reproduction. The idea of an animal, as Aristotle understands it, is the idea of an entity that preserves its form in part through its consciousness of its environment, and its resulting ability to respond to its environment in ways that serve to maintain its form. (Korsgaard 2020:17)

The way in which animals self-maintain is via representing the world around them and mentally guiding their behavior.[36] Animals are not just a different species of plant, rather they are a whole different kind of life. The fact that animals represent

[35] There are other types of life (bacteria, fungi, etc.) but the relevant distinction here is between purely nurative life (plants, bacteria, fungi, etc.), life that perceives and acts upon its representation of its surroundings ("animals" as I'm using the term), and then humans who are a unique form of animal life that also act self-consciously. Which is just to say, the classifications I offer here aren't meant to line up to those of the biological sciences which might cut these distinctions slightly differently because such scientists have different aims.

[36] Another relevant passages from Korsgaard: "[I]f the organism is [...] an animal or a person – the way that she constitutes herself is in part by having conscious states that track, at least roughly and defeasibly, what is good or bad for her in the functional sense." (Korsgaard 2014).

their environment and pursue self-maintenance in this way isn't just an additional power which animals possess over plants but rather fundamentally changes how animals live:

> But these are not just powers added, so to speak, on top of the animal's nutritive and reproductive life: they also change the way the animal carries out the tasks of nutrition and reproduction. The animal's capacity for perception and action determines the way it gets its food and ensures the existence of its offspring. (CKE 142)

The way in which animals self-maintain is via representing the world around them and mentally guiding their behavior. This is done mostly via an animal's instincts — the animal's in-built dispositions for how to react to perceived stimulus:

> The principles that govern an animal's movements as he guides himself through his environment—the principles that govern his reactions to his perceptions—are what we may call his instincts. [...] They determine what he does in response to what, what he does for the sake of what. (SC §5.5.5)

A spider's "organs, instincts, and natural activities" are all arranged toward the end of maintaining — "primarily through nutrition" — and also reproducing itself. (Korsgaard SC §2.2.1, §5.4.1). Reproduction is understood here to be a kind of self-preservation.[37] And a spider has a unique way of doing this different from a plant or a cockroach or an antelope. Spiders perceive their environment and react according to inbuilt rules that constitute the distinctive way in which a spider pursues the ends of self-maintenance and reproduction. A spider's instincts — like its organs, natural activities, and so on — are all arranged toward these two ends. Consequently, in

[37] In the case of plants or grass, reproduction seems more plausible as a kind of self-preservation. However, to view reproduction as a form of self-preservation in the case of humans who possess — as we shall see — more individuality than blades of grass. Which is to say, I find it questionable whether reproduction can be taken to be a form of self-preservation when the form of life under consideration possesses the further sense of individuality we will later discuss human agent's possessing. Therefore, for human agents I am less confident whether reproduction can be so easily counted as part of a human agent's self-preserving activities.

normal cases, we can say that when a spider acts from instinct, it acts from laws constituting its form. A spider's instincts and its form will only come apart if through some special circumstances — e.g. trauma — where the spider's instincts no longer direct it toward conducting its life in a spidery manner.

There are no self-aware spiders (for reasons yet to be discussed), but for illustration purposes let's imagine there were: a spider who did not act from first-order instinct, but instead a second-order instinct to be a spider.

> **A SECOND-ORDER SPIDER (SOS).** This is a spider in all physical respects, but it lacks the natural instincts of a spider. When it feels vibrations in its web, it doesn't instinctually respond by crawling toward them like a normal spider. However, this spider can represent the principles by which an ordinary spider reacts and — once we've additionally bestowed this spider with the in-built goal to be a spider — it can intentionally mimic the responses of a spider by doing what it thinks a spider would in its circumstances. Such a spider is acting from its representation of the laws that would make it a spider rather than just naturally having these as instincts. Nevertheless, such a strange spider does come to live a spider's life just with an additional level of complexity.[38]

[38] In some sense, this self-awareness means such a second-order spider lives a very different life than a normal spider. Perhaps it's better to say that such an organism lives a unique form of life that resembles an ordinary spider's life in certain respects but differs in a particular crucial respect (i.e. the manner it self-awaredly pursues spiderly activities rather than instinctually). Nevertheless, it makes a life for itself because by self-awaredly behaving as a spider would, it becomes self-maintaining just as a spider is self-maintaining. I ultimately think this raises some serious issues later on which I will return to address later. Let's call this the "Self-aware Non-Identity Problem" in that the issue is that the self-aware pursuit of a certain life isn't identical to living the life in a non-self-aware manner and set it aside for the time being.

A self-aware spider who acts from a second-order level instinct driving it to be a spider has a certain advantage in pursuing its self-maintenance over a normal spider. Consider a case wherein an animal's instincts do not align with their form. A story:

> Females of a certain species of Costa Rican wasp lay their eggs on the abdomens of a species of spiders called *Plesiometa argyra*. The larva lives off its host's equivalent to blood as the spider goes about its normal web building and insect-catching behavior. Then, after about two weeks, the larva injects the spider with a chemical "that makes it build a strange, new kind of web unlike anything it's built before. But this new web isn't for the spider: It's meant to support the cocoon that the wasp larva will build after finally killing and eating the spider." (Bates 2018)

In this instance, the spider's instincts result in actions which are not a product of the spider's form. A spider is not instantiated by spinning a web to support the cocoon of the wasp larva in its abdomen, even if the spider is acting from instinct. A normal spider is doomed in this case. A normal spider acts unreflectively from instinct. However, in our example of a self-aware spider, even if such a spider is tempted by the first-order instinct to spin the web the chemical in its brain compels it to build, whether it follows the direction of this instinct will depend on its second-order disposition toward living a spiderly life. A self-aware animal agent of this sort who operates from a second-order instinct enables it to better pursue its life when its first-order instinct goes wrong. What happens when its second-order instinct goes wrong? Well as I am imagining the second-order spider, it doesn't have any third-order or higher way to correct for that.

The spider possesses a second-order self-conception which governs how it behaves. It does so guided by this second-order disposition to constitute itself as a spider. Remember that a spider has activity as its form — that is, a certain collection

of actions which together add up to what it is to self-maintain in a spidery way. So a second-order spider is disposed to regulate whether it follows through on a first-order disposition or not depending upon whether that helps it behave in a way that constitutes living a self-maintaining life in a spidery way.[39]

---

[39] Second-order spiders illustrates an intermediate level between non-human animals and human beings. What about ordinary spiders? When we are thinking about lower-level non-human animals like spiders, neither the ends they pursue nor the principle it is guiding its movements are before their mind:

> At this level, [t]he animal is directing her movements and her movement are intentional movements—the movements have a purpose. In that sense the animal acts with a purpose, but at this stage there is no need to say that this purpose is somehow before the animal's mind. (Korsgaard 2006:108)

But what's precisely different about higher-order animals is their awareness of the ends for which they act:

> [A]n [intelligent] animal that can entertain his purposes before his mind, and perhaps even entertain thoughts about how to achieve those purposes, is exerting a [different sense of] control over his own movements than, say, the spider, and is therefore [an agent in a different sense.] [...] [I]t is at this level we become committed to keying the intentional description of the action to what is going on from the agent's own point of view. [...] This is a difference from the earlier stage: when we do describe the spider as "trying to get food," we don't care whether that's what the spider thinks she's doing. At the level of the spider, it is natural for the intentional description of the movement and the explanation of it to run together in this way. But once purposes are consciously entertained, the intentional description of the action must capture something about the way it seems to the agent. (Korsgaard 2006:109-10)

Once purposes are consciously entertained another change follows. Higher-order animals still behave according to their in-built instinctual principles, however because they are aware of the ends for which they act these animals can pursue these ends "intelligently". A dog both trained to bark at the front door at the signs of an approaching visitor and punished for stealing food from the dinner table, may come to bark at the front door not because he detects an approaching visitor but to draw his human owners away from the dinner table so he might steal the food he hungers for unimpeded. That is, given that he is aware of what his ends are, a higher-order animal can think instrumentally about means for accomplishing its ends and act from these thoughts intelligently in a way a spider or other lower-order non-human animal cannot.

Korsgaard credits Hume as being right about how deliberative decision-making works - albeit only for higher-order non-human animals (i.e. not human agents):

> According to Hume, the role of reasoning is to ascertain the relations between things. The only relation he thinks could conceivably be directly relevant to action is the causal relation (T 2.3.3,413–414). Knowledge of that relation can motivate us, but only if we have a pre-existing desire to attain or avoid one of the two objects thus related. For as Hume seems to picture it, our knowledge of the causal relationship functions hydraulically, providing a conduit by which motivational force passes from the desire for the end to the idea of taking the means, thus making the idea of taking the means desirable. (SC §4.1.1)

Korsgaard continues this same thought elsewhere:

This brings us to human agents. One natural way to think of human agents is analogous to how I described the self-aware spider. The spider possesses a second-order self-conception which governs how it behaves. It does so guided by this second order disposition to constitute itself as a spider. Similarly, we also operate in this manner. We work from what Korsgaard calls "practical identities". A practical identity is an identity I can come to possess by acting from the idea of how someone possessing that identity would act. This is essentially what my example of a second-

---

> And, moreover, if for whatever reason on this occasion I fail to come to desire taking the means, I am guilty of no irrationality or failure. It is merely that the typical psychological regularity failed to hold on this occasion. [...] This is how it is with intelligent but non-rational animals, and, if Hume were right, this is how it would be with us. (CA 64)

How an intelligent non-human animal responds to its affective states (i.e. its emotions and its instinctual or learned desires) are given to it by nature. "An animal's instincts tell it to hunt when it is hungry, flee when it is afraid, fight when it is threatened, and so on." (Korsgaard 1998:50).
A non-human higher-order animal may engage in instrumental thinking and thereby be inspired to desire the means to their ends and take those means, but this is not "rationality" but rather what Korsgaard terms mere "intelligence".

Importantly, these intelligent animals cannot deliberate about ends - only means to ends. For such animals, decision-making and deliberation is slave to passion just as Hume purported reason to be. Intelligent non-human animal's ends are determined by inclination rather than chosen by them. Non-human animals "are, in Harry Frankfurt's phrase, wanton: they act on the instinct or desire or emotion that comes uppermost." (Korsgaard 2006:103):

> Even in a case where the animal must choose between two purposes—say a male wants to mate a female but a larger male is coming and he wants to avoid a fight—the choice is made for him by the strength of his affective states. He has learned to fear the larger male more strongly than he desires to mate. The end that the animal pursues is determined for him by his desires and emotions. (Korsgaard 2006:110)

As such, animals do not choose their ends or actions in the sense that human agents do:

> There's [an] end—as in the case of [a non-human animal it's whatever end is set by which of its inclinations comes foremost]—[an] end or act that he's going to pursue or to do no matter what, and it rules him. And for him that end makes anything worth doing, anything at all, and that's a fact that is settled in advance of reflection. (SC §8.4.4)

Korsgaard further remarks, "our sense that there is something mechanical about him is not accidental. But he doesn't decide what is worth doing for the sake of what, because for him, that's already settled." (SC §8.4.4).

order spider was doing by regulating its behavior by what it needs to do in order to fulfill the conditions of a spidery way of life.

Human beings are animals, and so like spiders, our actions must be the result of mental guidance. This is because unlike plants, our form just so happens to be that of a living being who self-maintains their form of living by way of mental guidance. However, human agents differ from the other animals. The key here is that a human agent does not just represent the world and act reflexively from some dispositional instinct, a human agent is self-conscious of her own dispositional instincts and so is at every opportunity to act also faced with a choice of whether to do as her instinct bids. Which is to say, while a spider represents something in its environment and more-or-less reflexively responds by guiding its movements according to an instinctual disposition, a human agent is removed one level from these instinctual dispositions and is left free to choose according to a — let's call it — second-order or higher-order instinct regarding how she or he will respond to represented stimuli or first-order instinctual dispositions. What difference does that make? A significant one.

By governing my behavior by the idea of how a friend would behave, I act towards you as a friend would and thereby come to be a friend. I become an honest person by not telling lies even when I am tempted. I become reliable by keeping my promises and making myself available to others. I become a prudent agent by acting prudently. Moreover, there are identities which we may only think of descriptive which in fact have parallel "practical identity" versions. I can be a brother, or an American citizen, or a parent, or a scientist by job description all due to facts entirely unrelated to how I behave. However, there are "normative" versions of these

descriptive identities which are in fact a matter of how one behaves. We say things like “A parent puts the well-being of their children above their own” and “He’s your brother you have to help him if he’s in need.” It is of course true that not all parents put the well-being of their children above their own and that there is no causal descriptive necessity that someone must help their brother if he is in need, but when we say such things we are invoking the parallel “practical identity” versions of the concepts of parent or brother. A parent is a parent just by having biological offspring and a brother is just a brother by having a fellow male off-spring from the same parent, but these aren’t the notions of “parent” or “brother” at play when we talk in the way I’ve mentioned. When we talk in such ways, we are invoking the idea of standards which someone can or fail to live up to and by living up to they make themselves into a parent or a brother in this practical identity sense of being a parent or a brother.

And the notion of a practical identity is crucial to making sense of the idea of a fully self-determining agent or “free will”. This is the type of agency exemplified by human beings. Within the event-causal picture of the world there is no place for free will — at least not a libertarian conception of free will. A libertarian free will can only occur for living things from the teleological substance-causal conception of the world. A free will is a particular form of life and practical identity that self-aware agents can possess. It must be teleologically organized to maintain itself as a free will by engaging in free actions. This involves conceptualizing something teleologically as a self-reflexive teleological system. However, the notion of a free will cannot

occur at the level of plants or even lower animals — it requires self-conscious agency and the related notion of practical identity. Why?

A plant is in some sense autonomous. Its movements are the product of its teleological form and in that way separate from the event-causal explanation of the unfolding universe. In this sense all living things can be conceptualized as something that is free and autonomous. However, a plant is not responsible for the type of living thing or agent it is. Neither is any animal who simply operates according to merely first-order instinct. It's not even enough to be self-conscious and operate from practical identities like our hypothetical second-order spider. The choice of what sort of agent to be isn't part of their form of life. Even the second-order spider hasn't chosen its identity because the higher-order instinct to behave as a spider would isn't up to it. These agents are "free" in the sense that they are not causally determined by the event-causal chain, but they are not fully self-determined because they do not choose their own form.

However, if I govern my behavior by the practical identity of an agent qua agent — rather than a particular kind of agent — then in so doing I become an agent *but not any particular kind of agent*. If I were pre-determined to decide from a spidery practical identity or a cowardly practical identity or any other, then I couldn't be held fully responsible for being that sort of agent or the actions which flowed from it because I hadn't chosen to be that sort of agent. But what does it mean to operate from a second-order practical identity of agency itself rather than, say, our example of the second-order spider or someone acting from a practical identity of a friend or scientist or patriot or father? The most obvious illustration is the hypothetical

imperative (i.e. 'if we will an end, we must also will the necessary means to that end — or else we must abandon the end'). Thomas Hill Jr. provides a nice example of this:

> The paradigm of a person who offends against the imperative is the man who continues to declare himself for a goal, takes many steps toward it, and half hopes to achieve it even though he systematically refuses to take some means obviously necessary to reach the goal. For example, a man solemnly resolves to lose ten pounds of excess weight, buys smaller clothes, weighs himself each day hopefully, but rarely chooses the lighter meals that are required to do the job. (Hill 1992:20)

To have agency as a practical identity is, among other things, to not just operate from one's first-order instincts but to have as a second-order disposition toward satisfying constraints of agency itself such as the hypothetical imperative. Kantians maintain that the moral categorical imperative is also a feature of the practical identity of agency itself, but for now let's proceed as-if the hypothetical imperative were the only constraint built into the practical identity of an agent qua agent (as opposed to being any particular kind of agent).

While we do not choose the principles of agency nor the principles which make up our nature as reflective, free agents, we do choose the first-order instincts upon which we act so that we satisfy the conditions of agency like the hypothetical imperative. And so each reflective agent comes to have a unique form:

> Aristotle thought only God, or maybe the gods, had individual forms, but I've just claimed that it follows from his view that every human being has an individual form. Let me be clear about this. Your identity as a human animal, your human form, is given to you by nature, and you share it with the species. But the form of the human is precisely the form of the animal that must create its own form. [...] In other words, every person must make himself into a particular person. (SC §6.4.6)

And this is why we hold reflective, free agents responsible in a deeper way than other sorts of agents like plants and non-human animals:

> [E]very person must make himself into a particular person. [...] And it is because he makes himself into the particular person who he is that we hold him responsible for being who he is. [...] Sometimes you hear philosophers say that the idea of responsibility is incoherent, because we could not be responsible for what we do unless we are responsible for what we are, and we could not be responsible for what we are unless we created ourselves. I think it is true that we could not rightly be held responsible unless we created ourselves, but false that that makes the idea of responsibility incoherent. [...] [But] we are responsible because we have a form of identity that is constituted by our chosen actions. We are responsible for our actions not because they are our products but because they are us, because we are what we do. (SC §6.4.7)

The only principle which we do not ourselves pick and that is a part of our identity without our being responsible for it, is the second-order principle to act from only those first-order instincts which allow us to adhere to the principles of agency itself. We are not responsible for choosing to be free agents, but we are responsible for the particular sorts of free agents we freely choose to become. That is, we are responsible for the first-order practical identities / principles which come to define our individual character through our choices about what to do:

> If you choose to run in order to escape your predator, to stand your ground in order to protect your offspring, or to dance for the sheer joy of dancing, then those are your principles, your conception of what is worth doing for the sake of what. [...] [I]t is up to us to decide what justifies what, what counts as a reason for what, what is worth doing for the sake of what. We don't need to think of this, and in fact we shouldn't think of it, as a decision made prior to action: as often as not, it is a decision embodied in the action. It is because our actions are expressive of principles we ourselves have chosen, principles we have adopted as the laws of our own causality, that it makes sense for us to hold one another answerable in this way: to demand one another's reasons, and to take it, as we say, personally, when we hear what they are. (SC §6.4.3)

There is a sense in which all self-reflexive teleological systems are "free" agents. The actions of plants and non-human animals are conceptualized and explained

teleologically by the reasons they perform them given their place in the life-form of the agent. When we conceptualize these beings in this teleological way then we can come to see their movements as explained by appeal to the reasons for what they do instead of the mechanical, efficient causes of their movements we are conceptualizing them as free from the determinist, event-causal chain extending back to the original state of the universe. However, there is another sense in which the actions of these other living beings are not free. This is because these non-human forms of life do not choose their own self-reflexive teleological form. As Korsgaard points out: "[A] non-human animal's life is mapped out for it by its instincts; and any two members of a given species basically live the same sort of life (unless the differences are biologically fixed, as by age and gender, or by kinds as among bees)." (CA 142).

There is a deeper sense of freedom possessed solely by human agents. Human agents do not choose to be agents or to have the self-reflexive form of an agent, but they do choose the particular sort of agent they will be. In this sense, they have more control than non-human agents. Furthermore, it does not undermine the freedom of human agency to object that human agents haven't freely chosen their form as an agent qua agent. One might object, "But you haven't chosen the practical identity of an agent any more than the second-order spider has chosen the second-order instincts to regulate its behavior to become a spider". Which is to say, how is it that my choice of practical identities can be self-chosen? That must itself be the product of instinct or yet a further practical identity. It would seem then we cannot escape the need for an unchosen practical identity. This is true. But while on its face paradoxical, there is

one practical identity which — even if unchosen — does not compromise our freedom.

There is, however, a certain lacking of choice about what sort of agent to be which does not undermine our responsibility. The fact that you haven't chosen to be an agent doesn't make you any less responsible for your actions. For rather than complaining that any particular sort of action or form of agency was forced unchosen upon you, it is just being an agent that has been forced upon you. You are a free agent — even if you are not free to choose to be or not be an agent.

I've said this idea is paradoxical. Let me try and explain it one more time before moving on. To be bound to the second-order identity of a spider leads one to the actions of a spider and allowing one's movements to be explained in terms of the self-reflexive teleological form of life of a spider (assuming one had the physical set up of a spider allowing one to operate that way). But a human agent has the second-order identity of a free agent leading one to — for example — pursue consistency in action via the hypothetical imperative. Here one is making oneself not into a particular type of agent — i.e. not a spider agent or an antelope agent — but just an agent qua agent. By acting as an agent would, one makes oneself not into a particular sort of agent but just an agent simpliciter. Therefore one is free from any specific determination of what one does except those which follow from the necessary logic of agency like the hypothetical imperative. This is not a form of determination which

an agent can complain about as making them an unfree agent but rather simply what is necessary to make them an agent at all.[40]

§7. Our Contingent Practical Identities

A self-aware, second-order spider operates from a conception of itself as a spider. And in so doing, it becomes a spider. A similar thing is true of how we become someone's friend or a citizen or a father or even a being that cares about our own animal inclinations. A free agent is like this. It operates from a self-conception of itself as a free agent and in so acting becomes one. When it surveys its available movements it does what it sees as necessary not to be a spider (a particular kind of agent) or a prudent agent (another kind of agent) but to be an agent qua agent. What would a free agent do here?

---

[40] It seems to be an implication of my view that – for example – when someone fails to adhere to the principles of agency like the hypothetical imperative that they therefore not acting freely. This is a general problem that are often raised toward Kantian views. As Korsgaard frames the problem: "Because to the extent that an agent's [actions fail to adhere to the norms of agency] […] she is less of an agent, and to the extent that she is less of an agent, the source of her movements must be some force that is working in her or on her. […] Are we to say that the other characters become less responsible to the extent that there is less of agency about them, and more of the operation of some external or internal force?" (SC §8.5.1). I am inclined here to bite the bullet and say that failures of will aren't free actions. However, I don't see blameworthiness or praiseworthiness to depend on whether an action was free. In certain relationships (e.g. friendship, lovers) the other party to the relationship might be correct to hold an agent responsible even for their failures of will which – as failures of will — are thereby unfree and not chosen.

Korsgaard herself seems resistant to biting the bullet here and offers the following response to these sorts of worries: "Imagine a person I'll call Harriet, who is, in almost any formal sense you like, an autonomous person. […] In every formal legal and psychological sense we can think of, what Harriet does is *up to her*. Yet whenever she has to make any of the important decisions and choices of her life, the way that Harriet does that is to try to figure out what Emma thinks she should do, and then that's what she does. This is autonomous action and yet it is *defective* as autonomous action. Harriet is self-governed and yet she is not, for she allows herself to be governed by Emma. Harriet is heteronomous, not in the sense that her actions are caused by Emma rather than chosen by herself, but in the sense that she allows herself to be governed in her choices by a law outside of herself—by Emma's will." (SC §8.2.2). Myself I am not sure if this non-bullet biting response offered by Korsgaard succeeds.

What would an agent do? It would be motivated to act in accordance with the constitutive norms of agency. Here's one: the hypothetical imperative. Suppose I were aware that I had willed some means, then I would be motivated to take the necessary means to that end or else abandon the end. In the same way the self-aware spider has higher-level instincts that dictate that it does what it represents to itself a spider qua spider would do in these circumstances, and I have higher-level instincts that dictate that I do what I represent to myself that an agent qua agent would do. Such a being would not choose its instincts that drive it to seek being an agent, but nevertheless by doing what an agent qua agent would do, comes to be an agent. Not a spider agent or a prudent agent, but a pure agent governed by nothing other than the constitutive norms of agency themselves. And such an agent would be a free agent — fully self-determining.

A fully self-determining agent performs free acts in order to maintain itself as a fully self-determining agent. However this leaves us with a problem. The hypothetical imperative alone is not sufficient to guide action. I cannot be a free agent because there is nothing a pure free agent would do. Korsgaard's stated solution for this problem is obscurely stated at best, wrong at worst. Korsgaard proposes that this gives such free agents a reason to endorse contingent practical identities (Korsgaard SN 3.4.9).[41]

[41] Ultimately, I will argue that the practical identity of an agent qua agent is more substantive than just the hypothetical imperative. Not substantive enough to avoid the problem of contingent, non-moral ends. We still require our contingent practical identities to solve that problem just as I've argued. However, there are more restrictions on which contingent practical identities we can consistently hold than just the hypothetical imperative. Specifically, in order to account for diachronic agency, we must decide in a way that no other time-slice might reasonably reject. This gets us to prudence. But more than this. We must decide in a way that no other agent might reasonably reject.

Our problem is that it is not possible to just be an agent rather than a particular sort of agent. There is no way I can operate solely from the practical identity of being an agent and by acting become an agent. This is because there doesn't seem to be anything that a pure agent would do. The hypothetical imperative only constrains our choices of ends once we already have some.[42] Even – if in addition to the hypothetical imperative — prudence and the Kantian moral categorical imperative are taken to be further constraints built into the practical identity of agency itself, we still are not provided with much guidance about what to do.

Therefore, Korsgaard argues, the practical identity of agency itself provides reason for picking further contingent practical identities and thereby what sort of particular agent we will be. A creature operating from just the practical identity of agency itself would, first, pick several specific practical identities to adopt or endorse and make themselves into a particular kind of agent freely.

However this way of putting things can be misleading — *which is to say the way Korsgaard put it is misleading* (SC §3.4.9). It makes it sound as if I stand first as a pure free will and as such choose particular practical identities — in truth the historical way this occurs is the reverse. We grow up into adulthood already possessing a jumble of many contingent practical identities. And at some point we come to recognize that living by these practical identities is actually optional. Perhaps what enables this optionality is simply that from the perspective of this or that practical identity we can question, modify, or reject others. However, that isn't what

[42] That is to say, insofar as you continue to hold some end, you must also have certain other necessary means towards it as ends also.

Korsgaard thinks is going on here — or at least not all that is going on. What we come to recognize is that all of these practical identities are optional en masse. In fact, they must be if I am to see myself as something freer and more responsible than a mere non-human animal.

The perspective I come to occupy when I see all these practical identities as optional is that of being an agent qua agent. I might worry from here that everything turns out to be meaningless. That if all these practical identities don't have to be chosen, then I have no reason to choose any of them. However, I do have reason to choose them: I need them to be an agent. I cannot make myself into an agent qua agent unless I have some specific sort of agent I am making myself into, and so to make myself into an agent at all — to live by the practical identity of agency itself — I must make myself into a particular sort of agent by picking some specific further practical identities.

What this enables me to do is what I wanted in the first place, to re-endorse the practical identities I already was living by. In some sense what this line of reasoning endorses is just arbitrarily picking some to live by, but in truth we already have many which we are committed to and just stand in need of a way of endorsing from this higher perspective so we can feel comfortable in our endorsement of them. We don't need to arbitrarily pick contingent practical identities because we already began committed to many but stood in need of a way to endorse them as free agents rather than as just as causal givens by nature.[43]

[43] "According to Kant, then, to think thoughts about what you ought to do is at the same time to think thoughts about what you would do were you a fully self-determining being. And if it is possible for us to act as we would act if we were fully self-determining beings, then we are, for practical purposes, fully self-determining beings. This is why the content of the thoughts that move us can make a

Descartes is sometimes accused of pseudo-doubting all he knows to then just build back to everything he previously believed without ever "really" doubting it in the first place. This is meant as a criticism to claim that Descartes hasn't sincerely justified back to everything he believed previously, but built back to it after his "doubting" phase only because he was always aiming at re-establishing the beliefs he was previously merely pretending to doubt. I am claiming that something similar is happening here — only without the objectionable aspect. We start with many contingent practical identities which we come to realize are merely optional and — without giving them up for now — we seek out a way of justifying them from our perspective as free agents. This does not necessarily leave everything as it was before, some of our initial practical identities may have to be given up or modified in order to, for example, bring our lives in-line with the hypothetical imperative. However, by-and-large our initial contingent practical identities survive the reflective scrutiny that we subject them to once we back-up from them and evaluate them from the perspective of a free agent.[44] We realize that we are free from our starting-point practical identities and seek to find a way to keep them — perhaps with modification — from the practical identity we reach via reflection — i.e. the practical identity of free agency itself.[45]

---

difference to the degree of self-determination we exhibit when our movements are caused by our thoughts." (CA 12).

[44] A mafioso might have larger changes to make to his already adopted personal identities than someone raised and indoctrinated with more altruistic or liberal practical identities. Nevertheless, I suspect that for most of us finding a way to bring our already endorsed contingent practical identities in-line with moral requirements would not involve a complete revision or abandonment. Even in the case of the mafioso, many of his other already held practical identities could survive a more moral outlook on life even if his identity as a mafioso might not. How much change would be needed for any particular agent would clearly be a contingent matter.

[45] Sharon Street (2012:40-59) objects that there is no point of view of one's own agency qua one's own agency alone (no "practical identity" of oneself as a bare agent). Street argues that would involve the

suspension of all values and then to be left asking why one should take anything to matter at all from the point of view of oneself as a pure agent who does not yet value anything and stands in need of being convinced to value things and to act. Street objects further that talk of the point of view of a pure agent itself trades on a misunderstanding of the notions of "need" and "problem". Consider the following pair:

> "To be a parent, one needs to have children."
> "To be an agent, one needs to have reasons (i.e. have practical identities)."

The first is true of a parent, but if the agent isn't already a parent then it doesn't follow that the agent now has the problem of needing a child. Similarly, Street argues that while to be an agent you must have reasons (and so, practical identities) but if you take those away you aren't left facing the problem of needing reasons and practical identities. Street accuses Korsgaard of making this mistake in arguing that from the practical identity of agency we have reason to adopt further contingent practical identities:

> [Korsgaard is concerned with] the standpoint of an agent as such — i.e. the standpoint of a creature who is able to distance itself from its unreflective evaluative tendencies, and who needs an answer to the question "What should I do?". Such a creature, according to Korsgaard, "needs reasons," and therefore needs some normative conception or other of its identity to supply those reasons. […] [But this is] something akin to a "parent" who does not have children does not have a reason to have children; rather, he or she is not a parent at all. Similarly, an "agent" who doesn't take anything at all to be a reason does not have a reason to take something or other to be a reason; rather, "he" or "she" is not an agent at all. (Street 2012:50-1)

In what follows, I will be responding to Street's criticism here.

There is a way of conceiving of myself as a father normatively as Street envisions, but there's also another possible way. In this alternative conception: It's like owing it to my future unborn child to prepare the best life for him or her once he or she arrives. To set aside money for college now even before my child has been conceived and (perhaps) even before I've met their mother such that I am led to discriminate between possible mates on the basis of what sort of mother a potential mate would be to my child. That is, I might maintain a practical identity as a father in such a way that I see my future child as already real even if that child is in the future and so it is unlike — as Street argues — someone who cares about the norms of fatherhood but who doesn't thereby have reason to become a father. What I have in mind is someone who is already self-conceiving normatively as a father even if they don't yet have off-spring. Let's call this the *atemporal* version of a fatherhood normative practical identity versus Street's *temporally indexed* version of a fatherhood normative practical identity.

Do I have a reason to become a father when occupying the practical identity of fatherhood in this atemporal way instead of Street's temporally indexed manner? If I occupy the practical identity of fatherhood in this atemporal way, becoming a father isn't really a choice I make — it will already be assumed in my thinking. When I deliberate, what I am choosing between is whether and how much money to save for my child's college and who to mate with to raise my child with and so on. It's already taken for granted in my decision making that I am a father and that I have a child (who exists in the future) who I am acting for the sake of due to my practical identity as a father.

It's not the case that before having a child when I see a reason to put away savings for my future child or seek out a good partner to raise a child with together I am acting from merely a desire to be a father or the practical identity of "future father" or some such thing, but rather that I am already acting from the same practical identity of fatherhood that I do later in life if or when I actually have a child and see reason to take off work to help my child with his or her homework.

I take Korsgaard to be arguing that our reason to adopt a particular first-order practical identity is like *that*. I am deliberating while already taking for granted that I am a fully self-determining agent who will act from a first-order contingent practical identity, it isn't a question of whether to do so or not. In fact, the idea of it being a question I can decide is incoherent. Because that

§8. Conclusion

Here is the idea of human free agency introduced in this paper: Human agency has as its teleological self-reflexive form selecting particular contingent practical identities to endorse and live by — modified to be consistent with the background necessary and unchosen practical identity of agency — and thereby acting out the life of a free agent. Like my example of a second-order spider, a human agent has the second-order instinct to govern its actions according to the necessary rules of agency — like the hypothetical imperative — and so revise or abandon the lower level practical identities they are committed to by whether they cohere with this fundamental practical identity of being an agent. The second-order spider differs from the human given the content of what the second-order instinct involves between the two. The second-order spider has a second-order instinct to be a spider, whereas the human being has a second-order instinct to be a free agent. When the second-order spider succeeds, they become a spider; when a human succeeds, they become a free agent. Both the second-order spider and the second-order free agent is stuck with their second-order instincts, but one makes them into a spider whereas the other condemns them to being a free agent.

---

would be like facing the decision of whether or not to be an agent. But if I face that decision, I must already be an agent in the first place.

Do I have reason to adopt a first-order contingent practical identity? No, not really. I am here disagreeing with Korsgaard's manner of talking about such agents having a reason to choose further first-order contingent practical identities. Instead, a fully self-determining agent takes for granted in her decision-making that she has a first-order contingent practical identity (not one she's chosen and adopted yet, but one she will have and act from), and so her choice when she is deciding what to do is really *which* first-order practical identities to adopt and act from rather than *whether* to have a first-order contingent practical identity at all. Again: The choice she faces is *which* first-order practical identity will be her's, not *whether* to have one or not.

Paper 3: Diachronic Agency as Group Agency and Korsgaard's "Practical Identities"

§1. Introduction

In this paper I will introduce a theory of agency that treats individual agency as — paradoxically — a manifestation of group agency. On this strange view, the individual agent comes to exist at the same time as the group agent. My account starts with Parfitian idea of ourselves as a series of ephemeral time-slice deliberative decision-makers that occupy our body over our lifetime. How then does a unified, temporally extended agent arise out of these short-lived in-the-moment decision-makers? I argue that this happens when together these distinct deliberators deliberate from a common normative self-conception — e.g. as a father, as a US citizen, as a fully self-determining agent – thereby come to deciding together jointly what "we" as a team are going to do. By each time-slice doing its part in carrying out the joint decision they reached through joint reasoning, the time-slices perform a joint action and thereby come to constitute one persistent diachronic team agent. But such a view poses a worry: how can the basic form of agent arise from time-slice agents deliberating and acting together as a team? Doesn't that imply that the time-slice decision-makers making up the team were already agents before the formation of such a group agent?

To put my position in this paper less paradoxically: individual and group agency must be interdependent and holistically understood. There really can be no

individual agents without group agents of which they are apart and there can be no group agents without the individual agents that make them up via cooperative group action between them — both individual agents and group agents come into existence together. Individual agents on the view I will be defending are nothing more than spatiotemporal slices of a group agent. In this way individual agents relate to group agents like the individual pieces in the game of chess. There is no game of chess without the individual pieces of a pawn, knight, queen, king, and so on; but there are also no individual pieces without the game of chess to which they belong. The individual pieces of chess come to exist at the same time as the whole game of chess comes to exist and neither is metaphysically prior. Similarly, there are time slices of a temporally extended group diachronic agent who engage with one another in the right sorts of ways to generate a temporally extended diachronic agent, but there is also no such thing as time slices of a temporally extended diachronic agent unless there is a temporally extended diachronic agent in the first place. The temporally extended diachronic agent and the time slices of that temporally extended agent exist in an interdependency relationship wherein both come to exist together or not at all.

To argue for this view I begin with a problem embedded in the idea of diachronic agency and how the idea of group agency can help solve how me-today and me-tomorrow deciding together as a group agent can solve this problem of diachronic agency.

§2. Deciding for Tomorrow

There is a puzzle about what it means to decide for a later time. For example, what does it mean for me to decide today what to do tomorrow? Let's begin by exploring what it cannot mean. Deciding for tomorrow is not merely doing something today which controls my actions tomorrow. It isn't like there is a button in front of me which if I push that tomorrow I will robotically be controlled to carry out a certain plan of action. When I decide to do something today for tomorrow, there is still agency on the part of myself tomorrow that has to freely choose to carry out the decision.[46]

But this poses a problem for the standard model of future decision-making. It is often held that decisions for the future involve the creation of a mental state which — if it survives until the appropriate time — will control my actions then just like a decision for the present controls my actions now. But this shares a troubling similarity with the scenario where I push a button today which controls my movements tomorrow. It cuts out the decision-maker tomorrow in a way that decisions for the future do not.[47]

---

[46] Velleman can be found making the same point: "[O]ur future-directed decisions must not simply cause future movements of our bodies. If they did, our later selves would lack autonomy of their own, since they would find their limbs being moved by the decisions of earlier selves, as if through remote volitional control. We must exercise agential control over our own future behavior, but in a way that doesn't impair our own future agential control." (Velleman 1997:46-7).

[47] Velleman similarly also raises this concern about the standard view: "One might wonder whether there really is a problem here. If an agent forms an intention to do something in the future, and if he doesn't change his mind, then the intention will remain in place and eventually come into the hands of his future self. When the intention subsequently produces an action, the agent's future self will be acting of his own volition, since the intention producing the action will now be his. Yet whether an agent acts of his own volition, when governed by an intention remaining from the past, depends on the manner in which it remains and governs. If the intention is simply a lit fuse leading to action by some self-sustaining causal mechanism that's insensitive or resistant to the agent's ongoing deliberations, then it is not really a volition of his current self; it's just a slow-acting volition from his past." (Velleman 1997:47).

On the other hand, deciding for tomorrow cannot be just manipulation. Suppose I want to go for a run tomorrow and to force myself to decide to go for a 2-mile run tomorrow I give a friend a large sum of money to donate to a charity I hate if I don't go on that 2-mile run. This is very different from just deciding to go for a run tomorrow. Deciding today to go for a run tomorrow isn't like doing something today which coerces your future self to decide in a certain way tomorrow.[48]

One possibility is that the future decision-maker has some preference for "sticking to his guns" and the way he decides for the future is by deciding now about what to do tomorrow and thereby giving his future self an additional "stick to his guns" reason for doing as his past self decided. This also doesn't seem to be correct — although it is an improvement on the previous two models of deciding for the future in that it makes room for the agency of the future self. After all, is "sticking to my guns" itself a decision I've made for the future or just some inclination that I can reliably expect my future self to have and decide on the basis of?

The problem is that I am deciding today to do something tomorrow which I expect will be effective because of a desire my future self will happen to have. Suppose that my future self will have an odd compulsion or desire to do anything written on a certain scrap of paper. And so I write down something I'd like to do.

---

[48] Velleman, again, makes a similar complaint: "[But] employing these devices [e.g. relying on side-bets] would entail treating one's future selves as one treats separate people, since it would entail influencing their behavior indirectly, by modifying their incentives. If I offer you a large enough reward for following my directions, or threaten a large enough penalty for disregarding them, I put myself in a position to give you directions that will take effect without overriding your autonomy, but I do not thereby put myself in a position to decide what you are going to do. The ability to influence you by manipulating your expected pay-offs does not give me agential control over your behavior. Yet the control I would enjoy over my own future behavior via the devices under consideration would be no different." (Velleman). Velleman points out, though, at least on this model of deciding for tomorrow my agency tomorrow isn't crowded out. My decision tomorrow is coerced, but at least I have a decision to make tomorrow on this model.

Have I decided for tomorrow, or just decided for today on how to exercise my indirect control on my future self given what I know about my desires? The problem is that this would be a sort of control we could have over strangers as well. If I am aware that some stranger will do whatever I write on a certain board, then I've exercised the same sort of control in influencing his later decision through this means as I have for my own future self.

"But", one might object, "writing on a scrap of paper is different from making a decision. If my future self is inclined to do as I've previously decided, that's importantly different from my being inclined to obey whatever commands have been written on a scrap of paper." But is it so different? After all, what makes the mental act I perform today a "decision"? I say the words to myself in consciousness "Go to the dentist tomorrow" or perhaps I come to a certain particular sort of mental state directed toward going to the dentist tomorrow as the result of practical deliberation. But what makes these mental acts or states a decision for tomorrow? Well, they aren't a decision for tomorrow unless my future self has the appropriate disposition or inclination toward carrying them out. But then my future self's inclination can't be to carry out "prior decisions" since they don't become that until and unless I treat them a certain way. But if my inclination isn't to carry out past decisions, but to do whatever command in the past I've said to myself in conscious thought, then this doesn't seem all that different from having an inclination to obey any command written on a certain paper scrap.

But an alternative way of understanding what it means to decide for tomorrow is this: as myself today and myself tomorrow deciding *together*. The idea here is that

our diachronic, extended agency — the capacity to decide for later times — depends upon our deliberating, deciding, and acting jointly along *with* our future selves. In the next section, I'll begin to introduce my account of group agency to explain how it can be used to handle diachronic agency.

§3. Group Agency

Actions are performed by agents, but not always individual agents — sometimes agents act together. This is a distinction between two agents both individually happening to take a walk alongside one another and when they walk together. Philosophers speak of this phenomenon variously as "group" or "shared" or "joint" or "collective agency". Taken literally, they imply that in addition to the individual agents, there is a further collective agent to whom the action is ascribed to as well as some group mind possessing intentional states of its own.[49] In the view I will defend, it is literally true that group agents have intentions of their own which get created by individual agents deliberating together and carrying out the group intention as mere component parts of the overall group agent.

In an earlier paper I endorsed the Anscombian view of agency that "an individual action [...] is a rationally structured process-event [...] constituted by temporal phases—sub-actions—that are rationalized by the whole." (Schofield 2012:35). To causally explain a movement which constitutes an action is different from rationally justifying an action. A rational justification cites a broader process of

[49] On this point: "Does it ever make sense to say that a social group is the subject of, or just has, an intention, a belief, or an emotion? Or are such attributions of intentionality to groups merely metaphorical shorthand for referring to the attitudes of their members?" (Quinton 1975).

which the explanation is a sub-component; a rational justification does not cite a prior cause (not even a mental state like an intention which might seem to be what rationally justifies the action). Schofield (2012) argues that group action can also be conceived of on this model. A group action unfolds not only over time, but synchronically across the group's members. To answer the question, "Why is he doing that?" in a way that rationalizes an agent's participation in a group action, one must locate what the agent is doing not just as a sub-action in the diachronic unfolding of their individual action but also its synchronic role in a group activity. In executing a particular play in football, we rationalize the quarterback's movement not just in terms of his movement's place in the individualistic temporally unfolding process of his throwing the ball, but also synchronically in terms of how his throwing the ball plays a role in the group action of executing a particular play.

However, this is not an instance of each member of a group agent like a football team individually executing a Sellers-style 'we' intention but rather carrying out the singular 'we' intention of the group itself. In true group action, the end product of action is not able to be made sense of by rationalizing or tracing back what the individual agent does to their personal intentions or individual activities. It is only by appeal to the collective action of the agents as members of a group agent carrying out a group intention that we can make intelligible sense of what they are doing. That is, it must be the case that the rationalization of their actions ends entirely with the collective intention they share rather than with their own distinct individual intentions. It is the group agent's intention within which the individual agent's movements are to be justified and made sense of as action, there is not a manner of

making sense of the movements of the individuals as sub-actions in a larger unfolding action except as components in the group's action. The group is carrying out an action which each member does their part via their individual intentions.

Yes, the intentions of the group agent must be carried out via the intentions of the individual agents which make it up. But how is this possible without simply reducing to an action being the product of each individual agent's individual intentions? Suppose I decide to pick up the glass sitting before me and subsequently do so. I will do so with either my left or right hand, but I probably haven't decided on which hand to pick up the glass before me. I merely decided to pick up the glass and then my subconscious apparatus has carried out my intention in a way which runs through either my left or right hand. Which is just to point out that when I form an intention and execute it, many determinations regarding how I execute it are not made by me but simply done sub-personally. We can imagine something similar occurring in the case of group agents.

Alternatively, sometimes I have decided to go for a vacation to Dallas but have yet to decide on what flight to purchase a ticket for to get there. I might deliberate and come to a decision about which specific flight to buy a ticket for and then act on the intention to buy that specific ticket as a manner of carrying out my larger intention to vacation in Dallas without needing to think about the overall intention of whether or not to vacation in Dallas. In a similar way, individuals might deliberate about and act from specific "implementation" intentions regarding a larger intention possessed by the group.

At this point an example would help. Consider the following example from Korsgaard:

> Suppose you and I are related as student and teacher, and we are trying to schedule an appointment. ‘‘Stop by my office right after class,’’ I say, thinking that that will be convenient for me, and hoping that it will also be convenient for you. It isn’t, as it turns out. ‘‘I can’t,’’ you say, ‘‘I have another class right away.’’ So I have to make another proposal. It’s important to see why I do have to do this: it’s because having the meeting is something that we are going to do together. The time I suggested isn’t good for you, and therefore it isn’t good for us, and it follows from that that it isn’t after all good for me, and so I need to suggest another time. To perform a shared action, each of us has to adopt the other’s reasons as her own, that is, as normative considerations with a bearing on her own case. That’s why the fact that the time is not good for you means that it also is not good for me. So we both keep making suggestions and considering them until we find a time that’s good for both of us. (SC §9.4.6)

This we might take as our paradigm case of group agency. However, there are several features of this example which I think give the wrong impression about group agency. In this case the two agents are able to talk and deliberate together, that obviously isn’t the case if deciding with my future self is also an instance of group agency. Also, Korsgaard frames this in terms of — as she puts it — adopting “the other’s reasons as” one’s own. I think that is true insofar as it goes, but I think there is a better way to understand the idea that Korsgaard means to be getting at here and elsewhere which I will be exploring throughout this paper: the notion of deciding what to do from a ‘we’ rather than ‘I’ perspective. When the student and teacher reason from the perspective of “What are we going to do?” instead of each individually from the perspective of “What am I going to do?” the result is a joint intention possessed by their joint agency which each individual executes their part of rather than two coordinated separate individual intentions and actions. I will explain.

The result of "What am I going to do?" deliberative decision-making will be an action which makes sense from the frame of that individual agent's activity. Suppose such an agent is making an omelet and considers the question of what I am going to do? This reasoning will lead to decisions to go to the fridge, retrieving certain ingredients, turning on the stove, and so on. Likewise, if both the teacher and student are confronting one another from the individualistic perspective — the teacher aiming to fulfill their professional requirements of meeting with students and the student aiming to meet with her professor to review test answers — then they will come to meet at a time which can be embedded in the activities of each individual separately. This is made most clear if there were a manner of defecting. Even if they coordinate in this perspective they are treating one another's reasons as obstacles rather than as reasons. Whereas, if the choice of a time to meet is approached from the perspective of "What are we going to do?" this cannot be solved by defecting. Nor are the other person's reasons obstacles. More importantly, it cannot be made sense of from either agent's individual perspectives. If they adopt the "we" perspective primitively, then there is no individualistic making sense of why they do so. What they do can only be made sense of from the 'we' perspective. The actions of each agent to cooperate is only made sense of from the 'we' perspective and it is the action of the we which the individual agent's actions are sub-components.

One way in which we can illustrate the idea of team reasoning is to suppose we take two wholly self-interested agents and put them in a standard prisoner's dilemma scenario. A standard one-off prisoner's dilemma is as follows (where C

denotes cooperating — i.e. refusing to confess; D denotes defecting — i.e. confessing):

		Agent 2: C	Agent 2: D
Agent 1	C	4,4	0,3
	D	3,0	1,1

If both agents are acting rationally with no altruistic aims for their compatriot, the result will be D,D. That is, the outcome we can expect is that both of them will defect. But suppose, without modifying their desiritive set by adding some altruistic or cooperative desire, we simply ask of them that they frame their deliberative decision-making around a different question. Namely, not what is it that *I* am going to do but rather what is it that *we* going to do. Alternatively, we might imagine this happening spontaneously.[50]

Framed in this way, the outcome of their decision-making will be C,C — to both cooperate. Which is to say, the same devotion to self-interest and narrow pursuit

[50] "In some cases, such as military units, prison gangs and some ordinary workforces, [...] [i]f they are told to coordinate their actions to achieve a given goal, but the appropriate individual actions are left to the agents [...], acting on orders is team reasoning. These are cases of coerced agency. In other cases, it is more natural to think of agents not as coerced but as choosing to team reason because doing so is an optimal strategy in some embedding game. In yet other cases, team reasoning can arise as the spontaneous behaviour of agents who somehow or other come to 'identify themselves' with a group." (Bacharach 1998:3).

of the satisfaction of what they want in the moment that previously and expectedly led both to rationally defect, can without modification be parlayed to drive reasoning which now instead leads both to cooperate just so long as the deliberative question they take up is framed in the 'we' mode rather than the 'I' mode.

By thinking about what to do from the "we" mode, each agent comes to the conclusion for the two of them to pick C,C. Then each of them individually carries out this group action by picking C themselves. It is important that to rationally explain what greater action each agent is engaged in when they pick C cannot be explained by some broader action they themselves are carrying out. Just like before when an agent can be accurately described as "making an omelet" both when she is turning the stove on and then a moment later cracking an egg, each of these agents can be accurately described as engaging in the group action of choosing C,C together when each of them are individually choosing C as their answer.

We can extend this idea to also apply to intra-personal joint reasoning between myself at different times. Suppose I have a large pile of exams to grade and four days to do it. I have little else to do over these four days except for on the second day. On the second day I have to spend most of the day helping a friend move. Now, consider different possible ways in which I might split up my grading over these 4 days:

	Day 1	**Day 2**	**Day 3**	**Day 4**
Option A	25%	25%	25%	25%
Option B	0%	0%	0%	100%
Option C	100%	0%	0%	0%
Option D	50%	0%	50%	0%
Option E	30%	10%	30%	30%
Option F	0%	33%	33%	33%

To choose what to do in a "What are we going to do?" frame that includes myself on each of these four days, I deliberately seek a course of action that the agent on each day could agree to when deciding from the same 'we' frame from which I am deliberating. At first glance, it would seem like Option A is the course of action that myself on each day could agree upon if we were deciding on what to do from a "we" framework. However, since I mentioned that on day two I also have to help a friend move while on every other day have nothing else to do, perhaps it would also be possible for us to agree on Option E from a "we" framework. Whether that is possible would require us to get into the details of exactly how this "we" framework works, but I want to avoid working out the details of the account for now. The point is that, for example, Option B couldn't be chosen from the "we" perspective of myself on each of these four days. If I procrastinate on the first three days and end up leaving all of the grading I have to do to be completed on the fourth day, this can't be because I choose Option B from the "we" perspective and so as a diachronic agent carried out a decision to leave all my grading to the last day. Instead, what must have occurred is

that my diachronic agency failed to materialize. Afterall, that's exactly what procrastination is: a failure of agency.[51]

The model of diachronic agency that I've proposed here (one based upon my model of group agency) comes strikingly close to the model proposed by Natalie Gold in the Game Theory literature. To be an extended, diachronic agent requires one at each moment in one's agential life to treat one's reasons at other times as normative for oneself now and to reach a joint decision about what to do. Korsgaard's account here has influenced work in Game Theory where a similar view to her own has been proposed which I will now turn to.

§4. Decision Theory, Game Theory, and Team Reasoning

In the standard model within game theory, individuals are assumed to act rationally in pursuit of their own well-being. This has led to some notorious puzzles such as one-off prisoner's dilemmas (PD). In a one-off PD, while both individuals would be better off if they cooperated, what's individually rational for each agent is

[51] It might be asked what's wrong with Option C where I opt to do all the grading on Day 1 to get all this drudge work out of the way? On my account this constitutes a failure of diachronic agency just like procrastinating and leaving all the grading to be done on Day 4 as in Option B. For the result to be a joint decision by the team of time-slices, each time-slice deliberator's interests have to be respected. From the shared normative perspective of maximizing one's desires at the time, the time-slice on Day 1 has no more reason to agree to Option C than the time-slice on Day 4 has to agree to Option B. It is of course possible that the shared normative self-conception each time-slice is operating from to be that of a 'pro-active, go-getter' which gives each time-slice a reason to accomplish joint tasks as early as possible and that would give the time-slice on Day 1 a reason to agree to Option 1, but unless that self-conception is part of the shared practical identity which each time-slice's team reasoning is framed by then Option C would be a failure of agency just as much as Option B. It is also for each of the time-slices to adopt an internally "utilitarian" contingent practical identity and from that framework loading all the work on one time-slice for the overall benefit might make sense. It's obviously allowable for a community to freely adopt an arrangement which loads extra work on one member (who also accepts such an arrangement). Therefore there isn't anything preventing all of my time-slices from adopting a contingent practical identity which distributes burdens and benefits unequally due to some culturally instilled or otherwise adopted shared practical outlook among my time-slices.

that they defect. However, experimental testing reveals "that real individuals actually cooperate in unrepeated prisoner's dilemma almost half of the time (e.g. Sally 1995)." (Lecouteux 2018:3). The explanation within the standard model is to conclude that the participating individuals must possess altruistic preferences — "e.g. they are unhappy when payoffs are unequal in a PD, because they care about inequity for instance" (Lecouteux 2018:6).[52]

As I suggested earlier, suppose we take two wholly self-interested agents and put them in a standard prisoner's dilemma scenario. If both agents are acting rationally with no altruistic aims for their compatriot, the result will be D,D. That is, the outcome we can expect is that both of them will defect. But suppose, without modifying their desiritive set by adding some altruistic or cooperative desire, we simply ask of them that they frame their deliberative decision-making around not what is it that *I* am going to do but rather what is it that *we* going to do. Framed in this way, the outcome of their decision-making will be C,C — to both cooperate. This is provides us with an alternative to dealing with prisoner's dilemmas. I don't mean to be taking the empirical fact that people cooperate as evidence, I just mean to be suggesting that the game theory explanation for why they do may not be either irrationality on their part nor altruistic attitudes but rather because they've engaged with the problem using a 'team reasoning' attitude rather than an individualistic "I" attitude.

[52] It's important that shifting the frame from "What should I do?" to the team reasoning question of "What should we do?" doesn't achieve cooperation by simply giving the agent pro-social desires. Team reasoning is a different sort of solution for explaining cooperative behavior: "The common explanation of such unselfish behaviours is that players are individually rational (the theory of choice is therefore the right one), but that we must extend the definition of preferences to include unselfish motives" (Lecouteux 2018:3).

We find additional evidence that the 'team reasoning' approach is correct in other sorts of games where attributing altruistic attitudes wouldn't solve the problem Game Theory faces. Consider the Hi-Lo game:[53]

		Agent 2	
		H	*L*
Agent 1	*H*	5,5	0,0
	L	0,0	2,2

Surprisingly in Hi-Lo both HH and LL, either choice is rationally permissible according to the standard model. This is because I do not get a better payoff by switching from H to L or L to H unilaterally (i.e. both HH and LL are Nash equilibria). This is despite the fact that when it comes to actual people in Hi-Lo, almost everyone plays H and judges H to be uniquely the rational choice (Bacharach 2006:43; Lahno and Lahno 2018). As Bacharach remarks on Hi-Lo: "It is obvious that the only rational choice is to click on [H]. Yet oddly, game theory has no explanation of what makes [H]-choices rational." (Bacharach 2018:37).

[53] "You and another person have to choose whether to click on [H] or [L]. If you both click on [H] you will both receive £100, if you both click on [L] you will both receive £1, and if you click on different letters you will receive nothing. What should you do? It is obvious that the only rational choice is to click on [H]. Yet oddly, game theory has no explanation of what makes [H]-choices rational." (Bacharach 2018:37).

We can illustrate the issue raised by Hi-Lo for traditional game theory more plainly: If the other player chooses H, then I should choose H; if the other player chooses L, then I should choose L. However, I do not know what choice the other player will make and so I am stuck. The other player is in a symmetrical situation. Without either player having a reason to choose H while ignorant of what the other person will choose, there is no way for rational choice theory as it stands to prefer choice H while both players are reasoning individually. The other player wants the highest payoff just like I do, but the other player doesn't know what choice has the better payoff unless he already has some reason to expect what choice I will make - which neither he nor I have.

Again, the temptation is just to say that people are being irrational in preferring choosing A or acting from altruistic attitudes. However, altruistic attitudes don't work in this case. That is, the solution might seem to call for the same solution as we implemented for one-off PDs — i.e. assuming these players' preferences to include unselfish motives. However, surprisingly in these cases, even if we assume altruistic motives on the part of the players, this still does not enable the standard model to recognize HH as the correct outcome. The problem is that the altruistic choice will be H only if the other player picks H, whereas the altruistic choice will be L if the other player picks L. Since there is not reason to expect that the other player will choose H or L, adding altruistic preferences doesn't solve Hi-Lo in the way altruistic preferences can be added to solve Prisoner's Dilemma cases.

But, as we've seen, an alternative solution to Hi-Lo has been put forth within game theory. An alternative to positing altruistic preferences to solve the puzzle with

one-off PDs which just so happens to also solve the problem which crops up with Hi-Lo games. Again, this alternative approach was introduced into the literature under the label "team reasoning":

> The difference with social preferences approaches is that TR is neither selfish nor altruistic: it represents individuals as reasoning together about the achievement of common goals (Sugden 2011). It is not the preferences of the individuals that must be revised, but their theory of choice. TR may thus offer a more satisfying explanation of cooperation based on reciprocity rather than altruism: if P1 cooperates in a PD, it is probably not because she wants to maximize the payoff of P2, but because she expects P2 to cooperate too (so that they will be able to achieve together the 'good' outcome). Cooperation is not a matter of altruism (which is one-sided): it is the process of a group of individuals working together to promote a mutual advantage. Although I do not question the existence of altruistic motives or concerns for inequity aversion (which are included in the payoff function), the point here is that cooperation in a social dilemma is probably more a question of reciprocity and teamwork rather than of altruism. (Lecouteux 2018:32)

The basic idea of "team reasoning" within decision theory is that we might better model the rationality of cooperating in one-off PDs and for choosing H,H in Hi-Lo games if we take the decision-makers to be confronting the decision of "What will *we* do?" rather than "What will *I* do?"

The existence of framing itself is somewhat uncontroversial.[54] In Kahneman and Tversky's work on framing effects it is revealed how the same case framed in terms of lives "saved" versus those "dying" affects how individuals reason and decide. Seeing the decision as a choice through the concept of SAVING rather than

[54] "A frame is the set of concepts a person uses when thinking about the world. Frames are notorious because of Kahneman and Tversky's work on framing effects, where changing the description of a choice problem affects the choices that people make (Tversky and Kahneman, 1981). In their classic example, subjects were told that the US was threatened by a deadly disease, which is expected to kill 600 people, and asked to choose between two vaccination programs. Different groups of subjects received different presentations of the decision problem. One group received all the information in terms of how many of the 600 lives would be saved by each program, the other in terms of how many of the 600 would die. [...] [T]he presentation in terms of "saving" and "dying" influenced people's decisions." (Gold 2013:9).

the concept DYING involves the agent perceiving the decision differently similar to how an agent can perceive a line-drawing as either a duck or a rabbit depending on the concepts he or she is employing. Framing isn't a new philosophical idea introduced through *fiat* by decision theorists, rather: "Framing starts from the idea, familiar to philosophers, that seeing involves 'seeing as'." (Gold 2013:9).[55] Agents view the decision problem they face and the options they might take under a particular description. That is to say: "[I]n order to team reason, a player must have the concept "we" in her frame. [...] [S]eeing that a decision can be described as a problem for "us" is a necessary precondition for team reasoning." (Gold 2013:10).[56]

It has not escaped game theorists that this model of team reasoning can easily also be used to analyze intra-personal decisions as well as inter-personal decisions. In particular, Natalie Gold has argued in over twenty papers in the last decade that the "[p]roblems of self-control can have an analogous structure" to those which team reasoning are meant to solve (albeit, in the intra-personal case we are dealing with "an asynchronous prisoner's dilemma").[57] If every agent is asking the question "What

[55] Moreover, the role of "framing" in our decision making can be thought of as a necessity for us due to our "bounded cognition": "The standard agents of decision-theory use all the knowledge that they have, they always see their world under all of the infinite number descriptions available to them. However, real people are finite, so this is never a possibility for us. We have "bounded cognition". At any time, we will be using only a small subset of the concepts that could describe our situation." (Gold 2013:57).

[56] One problem area in the literature on team reasoning is this: work in game theory on TR has stalled on explaining the rationality of an agent taking up and employing TR rather than individualistic reasoning. Bacharach (1999, 2006) provides a potential evolutionary story on how team reasoners might have had an evolutionary advantage over individual reasoners, but Paternotte (2018) has thrown doubt on the plausibility of this solution. As we will see later, it's possible Korsgaard's arguments later in this paper might provide a potential answer to why people team reason and why it's rational to do so.

[57] As Gold puts it: "[In Decision Theory it] is conventional to analyse problems of dynamic choice as if, at each time t at which the person has to make a decision, that decision is made by a distinct transient agent, 'the person at time t'. Each transient agent is treated as an independent rational decisionmaker. [...] One objection to the decision theoretic account is that it provides a neat model of temptation at the expense of an impoverished notion of agency. Agency is entirely vested in the

should I-now do?" then we will have easy decisions to smoke, overdose on heroin, cheat on our spouses, etc. Decision theorists, Natalie Gold suggests, would benefit from the idea that diachronic agency between deliberators at different times can be made sense of in terms of team agency and 'team reasoning':

> The plan that the agent would most prefer to implement is to smoke a cigarette today and give up from tomorrow onwards (which is analogous to the case of pollution, where the most preferred outcome is to pollute whilst everyone else refrains) but that plan is not realizable because the transient agent of tomorrow will face the same preference structure and, hence, will not play her part. So we have a prisoner's dilemma with smoking equivalent to defect and refraining to cooperate. (Gold 2013:51)

The idea Gold is suggesting is that diachronic agency and self-control comes to be via individual time-slice agents adopting an intersubjective "we" perspective and solving their cooperation problem through team reasoning just as earlier game theorists have argued we solve cooperation problems like the prisoner's dilemma and the Hi-Lo game between strangers. It is easy to see the analogy between Natalie Gold's work and Korsgaard's views. In fact, Gold cites Korsgaard as an influence on her views. However, I think Gold's views face a serious difficulty. It is to that difficulty I turn to next because it will prove revealing about how we must understand Korsgaard's view and how it differs from what is being proposed by Gold.

---

transient agents; there is no notion of a self that extends over time. It is as if every transient agent asks 'What should I-now do?' (Natalie Gold 2013:50). The instruments for achieving self-control are limited to pre-commitments, taking actions that constrain the choices or alter the incentives that will be available to future selves; making a resolution in the hope that it will affect future behaviour is 'naïve' (Strotz, 1955-56). There is no sense of an extended agency over time, whereby earlier selves make plans for the person over time which influence their later selves because of their status of plans (as opposed to because the later self would have taken that action anyway). [...] [To answer this objection, I employ the] idea of multiple levels of agency has been articulated at the inter-personal level, where theories of team reasoning allow agents to ask 'What should we do?' and identify distinctive modes of reasoning used by people in teams (Bacharach, 2006; Sugden, 1993). I apply team reasoning at the intra-personal level, modelling the self as a team of transient agents over time." (Gold 2013:50).

§5. A Problem

Earlier I've introduced the idea of team reasoning using the ideas of reasoning from a "we" frame instead of an "I" frame. This is unproblematic when we are discussing two different agents reasoning together, but it falls apart when we try and apply it intra-personally as Gold attempts. The problem with Natalie Gold's account is how she models the self — as she puts it — "as a team of transient agents over time" engaged in intra-personal team reasoning (Gold 2013:50).

Korsgaard discusses the impossibility of deliberating from just the present point of view in terms of the impossibility of particularistic willing. Particularistic willing would be to make a choice about what to do "without taking it to have any other implications of any kind for any other occasion." (CA 123). To decide from the 'I' perspective of a deliberative time-slice would be an instance of particularistic willing because it is not deciding for any other time.

> If you have a particularistic will, you are not one person, but a series, a mere heap, of unrelated impulses. There is no difference between someone who has a particularistic will and someone who has no will at all. Particularistic willing lacks a subject, a person who is the cause of his actions. So particularistic willing isn't willing at all. (CA 124)

The worry here is that if it's just a series of time-slice deliberators acting from their strongest impulse of the moment, then there is no agent but rather just a body being moved around by whatever desirous mental state within that body that happens to win out by being strongest. That is not an agent moving herself, but a desire moving the body. In order for an agent to move her body there must be something over-and-above just the strongest impulse of the moment winning out. Rather, there must be a

temporally extended agent deciding what to do from a non-particularistic, team perspective. To see the problem, consider a story:

> Jeremy, a college student, settles down at his desk one evening to study for an examination. Finding himself a little too restless to concentrate, he decides to take a walk in the fresh air. His walk takes him past a nearby bookstore, where the sight of an enticing title draws him in to look at the book. Before he finds it, however, he meets his friend Neil, who invites him to join some of the other kids at the bar next door for a beer. Jeremy decides he can afford to have just one, and goes with Neil to the bar. While waiting for his beer, however, he finds that the noise gives him a headache, and he decides to return home without ever having the beer. He is now, however, in too much pain to study. So Jeremy doesn't study for his examination, hardly gets a walk, doesn't buy a book, and doesn't drink a beer. (CA 116-7)

Even if Jeremy's desires at each time end up leading him to successfully study and pass his exams, that wouldn't be a case of Jeremy acting but just the series of desires within him happening to push his body around in a way that brings about that result. For Jeremy to act he must show independence from his strongest impulse of the moment and to be able to make and pursue temporally extended courses of action. A series of time-slice decision-makers merely doing whatever they happen to desire most isn't an agent but rather a body being moved by whatever desire happens to win out in the moment. The team perspective from which we deliberate, decide, and act is actually the very perspective of a distinct "I" over-and-above our desires of the moment which is crucial to the notion of an agent acting rather than a body being moved by some mental state within her.

What about the 'we' frame? For example, deciding from a 'we' perspective involves a decision which applies now and has implications for whatever other times are included in the 'we'. Is this enough of an extended team perspective to get us away from mere particularistic willing and towards an exercise of temporally

extended agency? The thin 'we' frame as I think we naturally understand it — and as I take Gold to employ it — is a frame that is the simplest joint reasoning possible between two agents. In the case of group agency, 'we' reasoning is possible. This is because two agents already have reasons and interests which can be treated as equal in their joint 'we' deliberation about what to do. However, this is not the case with deliberative time-slices. A deliberative time-slice is not an agent nor does it have reasons or even a point of view from which to speak of its interests. In short: it isn't possible for deliberative time-slices to switch from 'I' deliberation to 'we' deliberation and thereby form a group agent because it isn't possible for deliberative time-slices to engage in 'I' reasoning in the first place.[58]

The same problem exists for how Korsgaard has described diachronic agency as being a form of group agency. Group agency, on her account, results from two agents taking one another's reasons as "public" between one another and reaching a joint decision about what to do from which each of them does their part. However, this does not seem possible when we apply it to the intra-personal case because momentary deliberative time-slices are not agents nor do they have reasons. It is not possible for these deliberative time-slices to take one another's reasons into account in joint reasoning because they don't have reasons.[59]

---

[58] Deliberating from the 'I' perspective of a deliberative time-slice shouldn't even phenomenologically involve the experience of an 'I' existing over and above the war of different desires within the agent. There is no separate agent deciding from a point of view that rises beyond the moment and exists over time distinct from just the body at the moment passively being moved by whatever its strongest desire turns out to win.

[59] Korsgaard says exactly this when speaking on why we cannot think of each of our deliberative time-slices as having reasons to cooperate and jointly reason with one another: "Perhaps it is natural to think of the present self as necessarily concerned with present satisfaction. But it is mistaken.
In order to make deliberative choices, your present self must identify with something from which you will derive your reasons" (CKE 372).

The perspective we occupy when we decide is removed from our impulses of the moment. It involves a 'freedom' from them. Where does this perspective come from? The idea of a 'we' frame on decision-making offers a suggestion. The idea is that there can be no 'I' from the perspective of momentary deliberators. These are simply wantons having their movements determined by the waring impulses within them. But by occupying a 'we' perspective in our decision-making, we create a perspective removed and above our momentary selves and a frame for decision-making in a way which is separate from the desires of the moment.

Korsgaard takes it as a given that the 'I do' must attach to an action just as Kant asserts that the 'I think' must attach to any thought. There must be a self doing the thinking and an agent doing the acting. But the 'I' is not some Cartesian pre-existing soul or consciousness that we discover by introspection. Rather, it is an attitude built into the pre-conditions of agency. Here, then, is a way of making sense of that thought. There are just momentary deliberators occupying my body in a series and which may just have their movement in that moment determined by the strongest impulse of the moment. In this case there is no persisting agent or agent at all — just a desire which wins out to control a movement for that moment. But by adopting something like the 'we' frame where one is deliberating about what to do jointly via negotiation with oneself at other times, a perspective comes into existence from which decisions can be made above and separately from the impulses of the moment. Moreover, an agent can come to exist which persists if this 'we' frame is taken up at other times such that group actions can be and are carried out.

It is by this ‘we’ attitude being occupied at different times, coming to joint decisions which then are carried out that group actions are performed and through this a group agent comes to exist.

The problem is, however, that the ‘we’ perspective cannot serve in this role. A ‘we’ perspective is only possible when moving from an ‘I’ perspective and to involve these different ‘I’ perspectives in team reasoning resulting in a group decision and action. However, as we have seen, this isn’t possible in the intra-personal case. In the next section I will introduce Korsgaard’s notion of “practical identities” and how they are meant to fill the role of the ‘we’ perspective in the case of intra-personal group agency.

§6. A Solution

The solution to our problem requires us to introduce Korsgaard’s notion of a “practical identity”. A practical identity is a frame for our practical deliberation like the “I” or “we” frame we’ve already discussed — in fact, we can think of both of these themselves as forms of practical identity. However, practical identities can also be more substantial than a simple “I” or “we”. A practical identity is a self-conception which frames one’s choices just like an “I” or “we” frame does.

Take the notions of fatherhood or American citizenship. These are both self-conceptions we can have of ourselves, but there are two forms this self-conception can take. To be a father is in a certain sense just a descriptive matter of whether one has genetic off-spring in the world and being an American citizen is just the descriptive matter of whether one is a naturally born or naturalized citizen. However,

both of these notions also possess normative counterparts. A father cares for the wellbeing of his children and raises his children, and an American citizen makes an effort to participate in the American democratic project by voting and cares for freedom of speech and other ideals which he or she is willing to volunteer to fight for. These are normative notions of these notions which an agent might succeed or fail to live up to — or might not care about at all even if they are genetically a parent or by birth an American citizen. And it is normative notions like these — e.g. fatherhood, American citizenship — which can serve as substantive frames for one's deliberation. If I deliberate about what to do as a father — in the normative sense — or an American citizen — in the normative sense — then I will reason my way to certain different conclusions than I would otherwise.

We can deliberate about what to do framed by these more substantial practical identities — e.g. fatherhood, citizenship — in just the same way we might deliberate from the "we" frame rather than an "I" frame. And, importantly, these substantial frames are also capable of generating diachronic agency just as the "we" frame could. If I deliberate from such a frame and it leads me to register to vote at one time and then go to vote at another, this establishes a persistent agent across different times. A group agent carries out an on-going action by my deciding on what to do at several times from the "As an American citizen" frame to vote and then at an earlier time registering to vote while at another time going to vote. This is one agent carrying out a single prolonged action because it is the product of myself at different times reasoning together via a shared normative self-conception.

Which is to say, a "we" frame is not special in creating a diachronic agent. What's important is simply that ourselves at different times occupy the same framing practical identity and thereby come to create a diachronic agent of that particular sort (e.g. father, citizen). In fact, concern for our impulses and desires of the moment are themselves just a practical identity which we may or may not take up to frame our deliberation. That is to say, one of substantive self-conceptions or "frames" from which we deliberate about what to do is that of ourselves as animals with interests and desires that matter in the moment:

> [Our actions usually embody] the thought that our own natural interests are worth conferring value on. [...] [I]n pursuing our own good and that of those whom we love, we are simply doing rationally what every animal does naturally. And many of the interests on which we confer value when we claim this status are natural interests that we share with our fellow creatures – our interest in freedom from suffering, in the satisfaction of our natural needs, in the enjoyment of our physical lives, and in the welfare of our offspring. I believe in pursuing our natural good in this way we confer value on our status as beings who have a good, beings who have interests. That is a status we share with the other animals[.] (Korsgaard 2021:36-37)

The diachronic agent is a product of each momentary deliberator deliberating from the same deliberative frame, but deliberating from a "we" frame is not enough on its own. The problem is that a "we" frame isn't enough to determine what to do since each momentary time-slice doesn't have reasons of their own to be combined in a "we" frame. However, if each time-slice proceeds also from a "As an animal with a natural good…" then the diachronic agent that is created by joint reasoning with different time slices occupying the same frame can reason in a way that ascribes oneself at each time as possessing interests.

§7. Conclusion

I began by assuming Parfit's picture of ourselves as bodies occupied by a series of ephemeral consciousnesses to be correct as opposed to any sort of continuing metaphysically continuous being. The relationship between the conscious time-slices occupying the body at one time and at another are as unrelated as much as total strangers. Parfit takes this to show that the time-slices which occupy our bodies are disconnected individuals. However, we have a model for how unrelated strangers can come to act together as an emerging agent: group agency. This is how I argue we come to be diachronic agents who arise over and above the individual ephemeral consciousnesses that occupy our body from moment to moment. The diachronic agent who emerges out of our conscious timeslices of the moment are the result of deliberating from a shared practical identity (like a 'we' or 'fatherhood' practical identity) which gives rise to a group agent that each momentary time-slice is only carrying out its implementation intention regarding. Obviously myself today and myself tomorrow cannot literally deliberate or coordinate together, but they can both deliberate about what 'we' are going to do from a joint practical identity. This, I argue, is both how different agents can come to be group agents and how the different ephemeral time-slice consciousnesses which occupy our body in a series come to be a diachronically extended group agent (i.e. we are all actually group agents).

Group agents come to act by different individuals taking up common practical identities and deliberating about what to do from those practical identities leading to coordinated action. This is how different agents come to generate an emergent group agent who acts itself as a separate entity. In the case of individual agents, time-slice

conscious deliberators take up common practical identities to generate a group agent over and above the ephemeral consciousnesses which occupy our bodies in a series. The time-slice decision-makers are not individual agents themselves but rather the group agent-at-a-moment and their individual reasons and interests which figure into the group decision-making is a product of what the group agent assigns to its parts. My desires at the moment get some weight in my deliberation because I adopt a practical identity of an extended animal who cares about our desires of the moment. Which is to say, the individual perspective and reasons of the agent at a moment who make up the group agent is itself a product of the group agent which the individual time-slices make up. Gold gets this wrong in treating the time-slices of agents as having default perspectives and reasons separate from the group agent they create by coordinating together. Korsgaard gets this wrong by talking loosely about how time-slice agents can create a persistent group agent by each time-slice respecting the reasons of one another in deliberation about what to do. Where both of these views go wrong is failing to recognize that the individual perspective, reasons, and values of the time-slices which make up the time-slices of the group agent are themselves a product of these time-slices being components of that larger group agent.

Lastly, in this paper I hope to have also explained how we might understand Korsgaard's notion of a "practical identity". It is a normative framework which can be deliberated from and via a common practical identity a group agent can be formed and — in particular — how this is how an individual agent can come to be formed as a type of group agent made up of ephemeral time-slices occupying our bodies in a series. The account of diachronic agency faces a number of objections which I cannot

take up here. Among them is the way in which my account makes us just as accountable to fulfilling the pursuits of our past selves which we intuitively take ourselves to be able to disregard in a way we are not able to disregard what matters to our future selves or others when we act. There are also concerns about to what extent joint deliberation and group agency is possible given how the temporal situation existing between myself today and myself tomorrow puts me in a dominating position with regards to my future self no matter how much I try and act as-if I were in dialogue with my future self. However, these are concerns I am confident can – in the end – be addressed in further work developing the account of diachronic agency which I have only begun the necessary work on here.

Bibliography

Amundson, Ron & Lauder, George V. (1994) “Function without purpose” *Biology and Philosophy* 9(4), pp. 443-69.
https://doi.org/10.1007/bf00850375

Bacharach, Michael (1999) “Interactive team reasoning: A contribution to the theory of co-operation” *Research in Economics* 53(2), pp. 117-47.
https://doi.org/10.1006/reec.1999.0188

Bacharach, Michael (2006) "The Hi-Lo Paradox" In Natalie Gold and Robert Sugden (eds.), *Beyond Individual Choice: Teams and Frames in Game Theory*, Princeton University Press, pp. 35-68.
https://doi.org/10.1515/9780691186313-006

Barandalla, Ana & Ridge, Michael (2011) “Function and Self-Constitution: How to make something of yourself without being all that you can be. A commentary on Christine Korsgaard’s The Constitution of Agency and Self-Constitution” *Analysis* 72(2), pp. 364–80.
https://doi.org/10.1093/analys/anq061

Bates, Mary (2018) “Meet 5 ‘zombie’ parasites that mind-control their hosts” *National Geographic*.
https://www.nationalgeographic.com/animals/article/141031-zombies-parasites-animals-science-halloween

Bishop, John (1989) *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.

Charness G. & Rabin M. (2002) “Understanding social preferences with simple tests” *The Quarterly Journal of Economics* 117(3), pp. 817-69.
https://doi.org/10.1162/003355302760193904

Chisholm, Roderick (1966) “Freedom and Action” In Keith Lehrer (ed.), *Freedom and Determinism*. Random House, pp. 31-47.

Clarke, Randolph (2000) “Incompatibilist (Nondeterministic) Theories of Free Will” In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.
https://plato.stanford.edu/entries/incompatibilism-theories/

Couch, Mark (2011) "Causal Role Theories of Functional Explanation" *The Internet Encyclopedia of Philosophy*.
https://iep.utm.edu/func-exp/

Cummins, Robert (1975) “Functional analysis” *Journal of Philosophy* 72 (November): pp. 741-64. doi:10.2307/2024640

Cummins, Robert (1983) *The Nature of Psychological Explanation*. MIT Press.

Davidson, Donald (1963) “Actions, Reasons, and Causes” *Journal of Philosophy*, 60(23), pp. 685-700.

Davidson, Donald (1973) “Freedom to act” In Ted Honderich (ed.), *Essays on Freedom of Action*. Routledge.

Davidson, Donald (1980) *Essays on Actions and Events*. Oxford University Press.

Davidson, Donald (1987) “Knowing One’s Own Mind” *Proceedings and Addresses of the American Philosophical Association* 61(3), pp. 441-58. doi:10.2307/3131782

D'Oro, Giuseppina & Sandis, Constantine (2013) *A Companion to the Philosophy of Action*. Wiley-Blackwell.

Fehr E. & Schmidt K.M. (1999) “A theory of fairness, competition, and cooperation” *The Quarterly Journal of Economics* 114(3), pp. 817-68. https://doi.org/10.1162/003355399556151

Foot, Phillipa (1972) “Morality as a System of Hypothetical Imperatives” *Philosophical Review* 81(3), pp. 305-16. doi:10.2307/2184328

Foot, Phillipa (2001) *Natural Goodness*. Oxford University Press.

Foot, Phillipa (2003) *Virtues and Vices*. Oxford University Press.

Frankfurt, Harry (1971) “Freedom of the Will and the Concept of a Person” *Journal of Philosophy* 68(1), pp. 5-20. doi:10.2307/2024717

Gold, Natalie (2013) “Team Reasoning, Framing, and Self-Control: An Aristotelian Account” In Neil Levy (ed.), *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199862580.003.0004

Griffiths, Paul (1993) “Functional Analysis and Proper Functions” *British Journal for the Philosophy of Science* 44(3), pp. 409-22. doi:10.1093/bjps/44.3.409

Hill, Thomas E., Jr. (1992) “The Hypothetical Imperative” *Dignity and Practical. Reason in Kant's Moral Theory*. Cornell University Press, pp. 17-37. doi:10.5840/monist198972320

Kerstein, Samuel J. (2002) *Kant’s Search for the Supreme Principle of Morality*. Cambridge University Press.

Korsgaard, Christine M. (1981) *The Standpoint of Practical Reason*. Doctoral dissertation, Harvard University.

Korsgaard, Christine M. (1996) *Creating the Kingdom of Ends*. Cambridge University Press.

Korsgaard, Christine M. (1996) *The Sources of Normativity* Cambridge University Press.

Korsgaard, Christine M. (1998) “Motivation, metaphysics, and the value of the self: A reply to Ginsborg, Guyer, and Schneewind” *Ethics* 109(1) pp. 49-66. https://dash.harvard.edu/handle/1/3122464

Korsgaard, Christine M. (2006) “Morality and the distinctiveness of human action” In Stephen Macedo & Josiah Ober (eds.), *Primates and Philosophers*. Princeton University Press. pp. 98-119. http://nrs.harvard.edu/urn-3:HUL.InstRepos:34257937

Korsgaard, Christine M. (2008) *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Korsgaard, Christine M. (2008) *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press.

Korsgaard, Christine M. (2014a) “The Normative Constitution of Agency” *Rational and Social Agency: The Philosophy of Michael Bratman*. Oxford University Press. pp. 190-214. https://doi.org/10.1093/acprof:oso/9780199794515.003.0009

Korsgaard, Christine M. (2014b) “On Having a Good” *Philosophy*, 89(3), pp. 405-29. doi:10.1017/S0031819114000102

Korsgaard, Christine M. (2018a) *Fellow Creatures: Our Obligations to the Other Animals*. Oxford University Press.

Korsgaard, Christine M. (2018b). “Animal Selves and the Good” In Timmons, Mark C. (ed.) *Oxford Studies in Normative Ethics*. Volume 8. Oxford University Press. pp. 55–77.

https://doi.org/10.1093/oso/9780198828310.003.0004

Korsgaard, Christine M. (2020) "The Origin of the Good and Our Animal Nature" *The Journal of Ethical Reflections* 1(2) pp. 7-28.
https://dash.harvard.edu/handle/1/37366069

Korsgaard, Christine M. (2021) “Valuing Our Humanity” In Richard Dean, and Oliver Sensen (eds), *Respect: Philosophical Essays*. Oxford University Press.
https://doi.org/10.1093/oso/9780198824930.003.0009

Lakatos, Imre (1970) “Falsification and the Methodology of Scientific Research Programmes” In Imre Lakatos and Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press, pp. 91-195.

Lecouteux, Guilhem (2018) "What do “we” want? Team reasoning, game theory, and unselfish behaviors" *Revue d'économie politique* 3(128), pp. 311-32.

Mele, Alfred (1992) *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

Mele, Alfred (2003) *Motivation and Agency*. Oxford University Press.

Millgram, Elijah (2009) *Hard Truths*. Wiley-Blackwell.

Millikan, Ruth Garrett (1984) *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, Ruth Garrett (1989a) “In defense of proper functions” *Philosophy of Science* 56 (June): pp. 288-302.
doi:10.1086/289488

Millikan, Ruth Garrett (1989b) “Biosemantics” *Journal of Philosophy* 86 (July): pp. 281-97.
https://doi.org/10.2307/2027123

Moran, Richard & Stone, Martin J. (2011) “Anscombe on expression of intention : an exegesis” In Anton Ford, Jennifer Hornsby & Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Harvard University Press.
https://doi.org/10.4159/harvard.9780674060913.c2

Nagel, Thomas (1970) *The Possibility of Altruism*. Oxford Clarendon Press.

Neander, Karen (1991) “Functions as Selected Effects: The Conceptual Analyst’s Defense” *Philosophy of Science* 58(2), pp. 168-84.
doi:10.1086/289610

Neander, Karen (1991) “The teleological notion of 'function'” *Australasian Journal of Philosophy* 69(4), pp. 454-68.
doi:10.1080/00048409112344881

Neander, Karen (1995) “Misrepresenting and malfunctioning” *Philosophical Studies* 79(2), pp. 109-41.
doi:10.1007/bf00989706

Parsons, Glenn & Carlson, Allen (2008) *Functional Beauty*. Oxford University Press.

Paternotte, Cédric (2018) "The Efficiency of Team Reasoning" *Revue d'économie politique* 128(3), pp. 447-68.

Peacocke, Christopher (1979) “Deviant Causal Chains” *Midwest Studies in Philosophy* 4(1), pp. 123-55.
doi:10.1111/j.1475-4975.1979.tb00375.x

Quinton, Anthony (1975) “Social Objects” *Proceedings of the Aristotelian Society* 75, pp. 1-27.
https://doi.org/10.1093/aristotelian/76.1.1

Sally D. (1995) “Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992” *Rationality and society* 7(1), pp. 58-92.
https://doi.org/10.1177/1043463195007001004

Schofield, Paul C. (2013) *The Commonwealth as Agent: Group Action, the Common Good, and the General Will*. Doctoral dissertation, Harvard University.
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11051217

Sehon, Scott (2012) “Action explanation and the free will debate: How incompatibilist arguments go wrong” *Philosophical Issues* 22(1), pp. 351-68.
doi:10.1111/j.1533-6077.2012.00234.x

Street, Sharon (2012) “Coming to terms with contingency : Humean constructivism about practical reason” In Jimmy Lenman & Yonatan Shemmer (eds.), *Constructivism in Practical Philosophy*. Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199609833.003.0003

Stoecker, Ralf (2003) “Climbers, Pigs and Wiggled Ears-The Problem of Waywardness in Action Theory” In Sven Walter & Heinz-Dieter Heckmann (eds.), *Physicalism and Mental Causation*. Imprint Academic. pp. 296-322.
doi:10.1057/9780230582972_16

Strotz, R. H. (1955-56) “Myopia and Inconsistency in Dynamic Utility Maximization” *The Review of Economic Studies* 23(3), pp. 165-180.
https://doi.org/10.2307/2295722

Sugden, R. (1993) “Thinking as a Team: Towards an Explanation of Nonselfish Behavior”. *Social Philosophy and Policy* 10(1), pp. 69-89. doi:10.1017/S0265052500004027

Sugden, R. (2011) “Mutual Advantage, Conventions and Team Reasoning,” *International Review of Economics* 58(1), pp. 9-20. https://doi.org/10.1007/s12232-011-0114-0

Taylor, Richard (1963) *Metaphysics*. Englewood Cliffs, N.J., Prentice-Hall.

Taylor, Richard (1966) “Action and Purpose” *Philosophy* 43(163), pp. 73-4.

Thompson, M. (2012) *Life and Action: Elementary Structures of Practice and Practical Thought*. Harvard University Press. https://doi.org/10.4159/9780674033962

Tversky, Amos & Kahneman, Daniel (1981) “The Framing of Decisions and the Psychology of Choice” *J Science* 211(4481), pp. 453-8. doi:10.1126/science.7455683

Velleman, J. David (1997). “Deciding how to decide” In Garrett Cullity & Berys Nigel Gaut (eds.), *Ethics and Practical Reason*. Oxford University Press. pp. 29-52.

Velleman, David (2000) *The Possibility of Practical Reason*. Oxford University Press.

Wilson, George M. (1989) *The Intentionality of Human Action*. 2nd Edition. Stanford University Press.

Wright, Larry (1973) “Functions” *Philosophical Review* 82(2), pp. 139-68. doi:10.2307/2183766

Wright, Larry (1976) *Teleological Explanations: An Etiological Analysis of Goals and Functions*. University of California Press.

Wouters, A.G. (1998) *Explanation Without A Cause*. Doctoral dissertation, Utrecht University.