Contemplating counterfactuals: On the connection between agency and metaphysical possibility

Sjur K. Dyrkolbotn [*1], Ragnhild H. Jordahl [†2], and Hannah A. Hansen [‡3]

[1]Durham Law School, Durham University, UK
[2]Department of Philosophy, University of Bergen, Norway
[3]Department of Information Science and Media Studies, University of Bergen, Norway

1 Introduction

We consider the connection between the metaphysics of modality and agency, focusing on how it can be captured in logics for reasoning about multi-agent systems. We argue that philosophical insights can be gained from looking to these formalisms and that they tend to come with implicit philosophical assumptions that we must consider if we are to understand their broader meaning.

Indeed, social structures that have been designed with the aid of formal tools have become increasingly relevant to social reality, for both real and artificial agents.[1] Hence philosophical assessment of logical tools appear especially relevant in this context. In addition, philosophy may offer interesting directions to pursue when developing these tools further. In this paper we argue for more research in this vein, and we point to the search for a formal representation of the so-called *dispositional theory* of modality as an interesting research challenge that seems particularly promising in this regard [Borghini and Williams, 2008, Vetter, 2011].

In Section 2 we give some background on metaphysical theories of modality in general and we argue in more detail for the claim that the connection between metaphysical modality and agency needs to be taken into account in order to

[*]s.k.dyrkolbotn@durham.ac.uk
[†]ragnhild.jordahl@gmail.com
[‡]hannaha.hansen@gmail.com

[1]The growing importance of the social web over the last 10-15 years serves as an obvious example of this development.

arrive at a proper understanding of both of these notions. We observe, in particular, that agency appears to feature crucially in important metaphysical arguments concerning possibility, while metaphysical possibility seems to be at play in important arguments concerning agency.

In Section 3 we give a brief introduction to the dispositional theory, emphasizing how it makes the connection between possibility, causation and agency clearer at the philosophical level. This adds further weight to the claim that providing a formal interpretation of this theory is an interesting research challenge. It seems to us that branching time temporal logics are particularly relevant in this regard, and in Section 4 we argue that variants of *alternating-time temporal logic* (ATL) [Alur et al., 2002] can serve as a fruitful starting point for such an inquiry. We also present some ideas for further technical developments that we think suggest themselves quite naturally on a dispositional reading of this formalism.

We mention that related work has already been carried out, giving a formal or semi-formal account of the dispositional theory [Vetter, 2013, Jacobs, 2010, Vetter, 2010]. But so far there has not, to the best of our knowledge, been any significant exchange of ideas between those working on this from a metaphysical angle and the computer scientists, logicians and epistemologists who have already been working on related formalisms for a long time, for instance in relation to the so called *social software* paradigm [Parikh, 2001]. We conclude in Section 5 by suggesting that the relationships between related formalisms should be considered further and, moreover, that making the connection between metaphysical modality and agency explicit can help hope to shed new light on a number of well-known issues, both from philosophy and the theory of multi-agent systems.

2 Why metaphysical modality?

> It is spirit to ask about two things. (1) Is what is being said possible? (2) Am I able to do it? It is to lack spirit to ask about two things: (1) Did it actually happen? (2) Has my neighbor done it; has he actually done it? (Søren Kierkegaard)

When philosophers speak of metaphysical modalities or metaphysical possibility, they refer to a notion of modality that is wider than the possibilities that the physical forces, natural laws or statistical evidence of our world dictates, but narrower than "everything thinkable". Everything that is imaginable or thinkable is not the same as what is metaphysically possible - we can see a division here, an important one, between what is seen as metaphysically possible and the metaphysically impossible. This is the notion of possibility that will be discussed in this paper, and the term "metaphysical" is used to make this distinction clear.

One of the main controversies in contemporary work on metaphysical modality arises from the tension between the theories of Lewis and Kripke respectively [Kripke, 1981, Kripke, 2005, Lewis, 1986, Lewis, 1971]. Both Lewis and Kripke

build on the account given by Leibniz [Leibniz, 1998], who held that something is possible if and only it is true in some possible world, and necessary if and only it is true in all of them.

Both of these theories embrace realism with respect to metaphysical modality – that possibilities are existing entities in some sense – but when we speak of Lewis' theory, we can refer to this as a *possible worlds realism*, as it relies on an ontology which posits the existence of concretely existing possible worlds, completely separated from our own, Kripke's theory is based on an *actualistic* understanding of possible worlds; what actually exists is taken to be that which is part of our world, and all that is possible must, in principle, originate from this actuality.

It is commonly accepted that a powerful argument can be made against Lewis' theory by considering *identity* and *de re* modal claims, that is, modal claims about a particular existing object. How can it be, for instance, that something which is possible for *me* is witnessed by the existence of some other world, all the while I myself am part of this one? Recall that Lewis' worlds are fully real, fully existing entities, so my existence here prevents me from existing elsewhere, the physical entity that is me cannot be two places at once. Lewis answers by saying that what is possible for me is witnessed by something which obtains in some possible world for someone who is not me, but is very much like me, namely my *counterpart* [Lewis, 1971]. This counterpart relation is considerably more vague than the identity relation, and opens for consequences that might be seen as problematic.

The notion of an essence has been particularly crucial in the philosophical debate on this point. These are fundamental properties of objects, features that bestow upon them their identity, serving as defining characteristics of what they are. In the case of agents with cognitive powers of reflection and contemplation essences are particularly important as they pertain to the crucial question of *personal identity*: what makes me who I am, how do I identify myself as an autonomous being in a complex system? One view among philosophers is that essences are "moderately tolerant" to change, see e.g., [Bricker, 2008] which argues that this view lets us preserve some meaningful intuitions without losing a meaningful notion of identity altogether. He argues, for instance, that we should view *some* difference in our physical origin as possible metaphysical options, but that this needs to be restricted, e.g., so that it is possible that I could have had one different parent, but that both could not have been different.[2]

To us, such an imprecise and permissive understanding of the notion of an essence is not convincing. Following Kripke, we agree that an object's origin – like me having the exact parents I in fact have – is part of an objects non-trivial essence, and thus something that individuate objects and makes it possible to think of an object existing in several worlds, with very different properties. This, moreover, is a stipulation that is not only metaphysically well argued in the work of Kripke and others. It also serves to make metaphysics relevant

[2]Since then I would no longer be myself; I would no longer have a well-defined identity modulo this counterfactual, thus making it *metaphysically irrelevant*.

to theories of agency and interaction; we should not consider scenarios which differ from actuality to the extent that identity cannot be held to have been preserved. This, however, is not a descriptive fact about agents, arising from the fact that such scenarios are unthinkable – clearly they are not – but rather a normative stipulation we should make, arising from the fact that such scenarios are *impossible*. It follows from our understanding of metaphysics that it would be *irrational* for an agent to contemplate such possibilities, for the simple reason that they are not real.

For Lewis and other counterpart theorist, this conclusion is hard to reach, since the assumption that possible worlds are real and causally unrelated to the actual world naturally challenges such a more restrictive view of essence and identity. For Lewis, rather, the problem becomes that of accounting for the epistemic access we seem to have to possibilities, all the while they are witnessed by completely separated alternative worlds. Kripke, on the other hand, does not seem to run into problems in this regard, since for him possible worlds are merely stipulated — they are an abstraction, and because of this we are guaranteed that they will contain precisely those objects we want them to contain, for instance me, but with different properties than in the actual world. This focus on possible worlds as stipulations makes the metaphysical theory tighter and more relevant as a limiting theory with possibly interesting consequences for theories of rationality, knowledge and multi-agent interaction. It rids us of the problems concerning identifying objects across worlds, as identity becomes seen as a *given*, and not a property that must be established by looking to properties of worlds.

The counterpart theory is also held by many to be an affront to our intuitive understanding of modality, and particularly with respect to intuitions about agency. For instance, we seem seem to be egocentric when it comes to questions concerning our own possibilities or when we contemplate counterfactual situations. We are wondering about *ourselves*: In a famous thought experiment [Kripke, 1981], Kripke makes this point by considering the possibility that Humphrey won the 1968 US presidential election. Why exactly would Humphrey care if someone very much like him won the election? Surely, when contemplating the possibility of victory, Humphrey is thinking about *himself*?

Some of the problems in the philosophical debate concerning possible worlds seems to stem from the metaphor itself — that taking the metaphor too far has created both problems and misunderstandings that might have been avoided if one could explain modality in a way that doesn't take possible worlds as a primitive notion. This, connected with a wish to create a satisfactory actualistic account of modality is also a motivation for leaving the possible worlds behind a bit, and rather focus on *this world*. We think the dispositional account of modality, which we describe in more detail in Section 3, can be part of the solution here, as this theory firmly roots modality in this world.

Another motivation comes from considering the vast landscape of different actualistic accounts of modality. The genuine realism has the advantage of being mainly Lewis' thinking, and as a result of this it is a much unified theory. Actualistic realism, on the other hand, consists of several different ways of ex-

plaining what a possible world is — a set of states of affairs, possible histories of the world, etc. But the dispositional theory does not need to meander on this point as there is no need to specify what a possible world is at all. Possibility, rather, is seen as an actual property of our world and our existence, and most expedient in our way of thinking, not some far fetched idea related to some alien entities.

Kripke's argument in favor of actualism, and the question of identity across possible worlds more generally, seems to owe much of its significance from considerations rooted in agency. Notice, for instance, that modal agency, involving an agent contemplating the possible, is the performative core of the Humphrey thought experiment. More generally, whenever a modal claim becomes pressing in real life, this is invariably due to some agent engaging in modal reflection.[3] Moreover, when doing so, the agent is invariably embedded in structures that are present in physical and social reality, and his thoughts may in turn give rise to actions that can *change* these structures. So if we take the earlier mentioned "egoism" in our modal thinking as a starting point, we can also move further to the contemplation of ones own possibilities in situations that arise — i.e. as backgrounds for choices, not only as a retrospective tool focusing on what might have happened. In this paper we turn the focus to the contemplation of what can, will or may come. We want to argue that modality matters — that the modal structure both of the world and of our thinking about the world has an impact on how we ground our choices, and that it therefore plays an important role in our rationality.

3 The dispositional theory

> In asking with regard to my own actuality, I am asking about its possibility, except that this possibility is not esthetically and intellectually disinterested but is a thought-actuality that is related to my own personal actuality – namely that I am able to carry it out. The how of the truth is precisely the truth. (Søren Kierkegaard)

On the dispositional account, the possible is determined by dispositions found in the actual world; we remain rooted in this world, and we describe modality as something that is *present* and *real* (e.g., not a phenomenon arising simply from the way we tend to use our language). To say that something is

[3] That is not to say that modal agency subsumes or is constitutive of metaphysical possibility; this would involve excluding many possibilities that are often included in a metaphysical account, such as the possibility of a world with no agents (some may want their metaphysical theory to exclude this, but we prefer to remain agnostic about it). We are not, in particular, suggesting any kind of fictionalism about metaphysical possibilities, and the point we are making is not subsumed by previous work in this vein, as that of [Rosen, 1990, Rosen, 1995].

While agency should also be considered by such theories, their primary concern is with how possible worlds are to be made sense of, and how they come to be. This is not our topic; our argument is that *regardless* of what possible states of affairs are, it appears that how we *interact* with these in our social lives is relevant, also to the formulation of an appropriate metaphysical theory of possibility.

possible means that there is some actual disposition for which this possibility — this possible state of affairs — is its manifestation. The (possible) manifestations can serve to characterize and individuate dispositions, but as dispositions themselves are actual, *they* determine what is metaphysically possible – what could possibly manifest – not the other way around. Then we need not rely on possible worlds (real or metaphorical) as a primitive philosophical notion. Possible states of affairs can still be modeled formally as points in a directed graph – a powerful tool in modal logics – but according to the dispositional account this does not imply any commitments regarding possible worlds, not even to their existence. Rather, possible states of affairs can be *traced back* to their origin in actuality, and while they have rich internal structure, this structure arises from how they could have come about, so that the discourse of possible worlds can remain entirely metaphorical without challenging the reality of metaphysical modalities.[4] If we "reduce" the possible worlds to this formal logic tool, and see them as that only, and not some important metaphysical entity, we hope to avoid the problems that this terminology has created in the past.

It is important to emphasize that dispositions always trace back to properties of objects present in the world here and now. New dispositions do not spontaneously appear along any (counterfactual) future time-lines, and all possibilities result from the possible manifestations of existing dispositions. Still, *higher-order* dispositions might need to be considered, i.e., dispositions that are merely possible and arise from manifestations of dispositions that are always closer – in a chain of possible manifestations – to dispositions existing in the actual world, see [Borghini and Williams, 2008]. At the present moment this will not be the center of attention, as it seems important to firmly establish a proper framework before considering these more unlikely or far fetched possibilities.

The actual manifestations of the dispositions is something that might or might not come about, and objects tend to have many dispositions that will never materialize. Think of the glass that has the dispositional property of being fragile — this means that the glass will break if struck with sufficient force, but this disposition to break might very well never become actual. But even if the dispositions are never manifested, the existence of dispositional properties is enough to account for the possibility that the glass *might* break or that it *could have been* broken.

The connection between agency and dispositions can be elucidated by considering the term *powers*. It is used in the philosophy of causation, often as a synonym for dispositions [Mumford and Anjum, 2011b], but also in the philosophy of agency, where it has a different, but related, meaning [van Inwagen, 1983]. Roughly speaking, a power can be seen as a disposition involving agency by way of pointing to an ability that an agent has to bring about an outcome. In the example above, one might say of the glass that it is disposed to break, but one might also say of an agent that he has the power to break it. It seems wrong, however, to say that he is disposed to do so, simply because he can.

We want to stress this distinction because it is useful for a dispositional

[4] We point to [Vetter, 2011] for a survey of recent work on dispositions and possibility

theory of possibility. If someone claims "it is possible for me to break the glass", it seems that the disposition of the glass to break if he hits it is no longer a sufficient truthmaker. What if, for instance, we consider a world where this person does not exist, or he is necessarily prevented from hitting the glass for some other reason? In this case, it seems natural to also make reference to his power to hit the glass, not only the dispositional fact that it might break if he does so. So if we focus on the agents and their contemplation on how to bring about some result, it actually seems like the powers term is the most important one, as it is this term that will denote what it is possible for the agent to achieve in a given situation. However these achievements are limited not just by what actions the agent can perform, but also by the dispositions of the object that the agent interacts with, and maybe also by underlying dispositions in the agent. So even if we reserve the term *powers* for the agents (or the conscious components of our model), these will still be closely interrelated with objects dispositional properties and the different dispositions stemming from these.

In the Humphrey thought experiment, Humphrey knew he lost the election in 1968, but he was still free to contemplate the possibility of a different outcome. By contemplating this possibility, he was engaging in a form of agency, and while this agency was certainly related to his actions in the actual world (or at least to his attitudes towards those actions), this does not appear to be a form of agency that we can easily reduce to other forms. For instance, it does not seem possible to readily account for it in terms of causal decision theory, which considers agency and rational choice in the light of philosophical accounts of causality, see e.g., [Joyce, 1999]. Indeed, it is too late for that; Humphrey has already lost, so for an account centered on utility-maximizing, his thoughts about winning, in hindsight, are simply *irrelevant*. However, as anyone who has ever entertained such thoughts would surely agree, this is a gross oversimplification. In particular, the judgment of irrelevance in this case is based on an understanding of possibility that is too narrow. These thoughts matter to Humphrey, and they might come to influence his future course of action, particularly with regards to what his *new* goals will be, and how he will go about trying to achieve those.

This is the case even if it is not easy for him to see, given the information available to him currently, how exactly these thoughts can contribute to utility-maximizing behavior and optimization of future choices. Still, there might be whole range of non-obvious courses of future events that will make them crucially important, perhaps only after Humphrey himself revises his goals and based on a new understanding about the meaning of the events that have taken place. Also, it might perhaps be contingent on a number of other possibilities that may be more or less realistic, thus representing contemplation about a possibility that does not arise from contemplating the consequences of particular choices, but is more directed at establishing whether there *could be* a sequence of events that would lead to its manifestation.

We remark that recent work in psychology flags a similar distinction between different forms of counterfactual thinking and emphasizes the way in which even the loser forms, not associated with concrete choices or easily identi-

fiable causal chains, also serve an important functions in sound human reasoning [Epstude and Roese, 2008]. On the other hand, it is also important to remember that *all* kinds of counterfacutal reasoning can be potentially harmful, not just those associated with more distant possibilities. Entertaining painful and debilitating regret, for instance, should often be designated as *irrational* even if they are concrete and easily traceable to particular choices that did not have the intended utility-maximizing consequences. Moreover, while it seems clear that instrumental rationality and causal decision theory does not cover all forms of rational counterfacutal contemplation, it also appears that they can sometimes be *too* permissive, unable to provide appropriate restrictions on the class of scenarios that we should consider.

For instance, think of the case of a gambler who bets heads in a high stakes wager but loses due to the coin landing tails. He might occupy his mind with regret and distress based on the reasoning that *if* he had bet tails he *would have* won. Hence he might feel entitled to conclude, as per subjunctive conditional and modus ponens, that he has in fact acted stupidly and lost as a result. In turn this might even motivate future choices, such as betting tails the next time, or in more severe cases (but certainly not uncommon), turning to lucky charms and rituals to improve the chances of success.

- That the gambler's choice could in any way influence which way the coin was going to land.
- That rubbing lucky charms or triggering any other actual dispositions could in any way be influential in this regard, or result in more knowledge on part of the gambler.

Recognizing this, we also come to recognize that the epistemological thought-experiments which rely on entertaining such possibilities are in fact mute. Both the independence of choice and outcome as well as the impossibility of knowing how the coin will land become metaphysical necessities (as opposed to mere physical or statistical facts), thus revealing that any perceived problems associated with this kind of contemplation are simply *unreal*. It is not true that the gambler *could have won* even if, according to a strictly deterministic theory, it would certainly appear to be true that he would have won, had he made a different choice. But even for a deterministic theory this is a metaphysically irrelevant observation which warrants no further attention. Indeed, our metaphysics here points to a simple fact about agency, namely that mere possibility of outcome X does not imply the possibility for an agent to *ensure* that X obtains. The truth that it was possible for the agent to win does not imply the truth of the claim that it was also possible for the agent to make a choice so that he would win. Treating these as metaphysically on par with one another is simply not appropriate, at least not on the dispositional account.

We here see how metaphysical considerations can provide a more subtle view on possibility that influence our theories of agency. To elucidate further on such connections, and particularly the consequences of the dispositional account, we should turn to formal models.

The interrelated nature of powers and dispositions is further underlined by the observation that mathematically speaking, the formal frameworks used in [Jacobs, 2010, Vetter, 2010] to study objects and their dispositions are strikingly similar to logics used to study agents and their actions in the theory of multi-agent systems. This, in particular, is the starting point for our technical project, which aims to give an account of the dispositional theory, as well as the connection to agency, by means of multi-agent logics.

4 Agency and metaphysical possibility in formal logics

There is a vast landscape of formal logics that involve agency and possibility, and increasingly, these notions are also considered together, especially in logics for modeling interaction in a multi-agent system, see [Wooldridge, 2009, van Benthem, 2011]. Here we will rely on *multi-modal* logics, allowing us to study interactions between a modality representing metaphysical possibility, and another, distinct modality, which can be used for talking about agency involving reflection concerning such possibilities.[5]

In this regard, it seems natural to focus attention on logics that are based on a branching time notion of possibility. Such logics have attracted much interest, both in philosophy and AI, and they are particularly interesting because they have been extended in various ways by adding modal operators specifically directed at modeling agency. We point to [Belnap and Perloff, 1988, Horty and Belnap, 1995, Alur et al., 2002, van der Hoek and Wooldridge, 2003, Ågotnes et al., 2009, Broersen, 2011b] for a collection of work on such formalisms that seem relevant for the study of dispositional possibility.

To see how branching time formalisms can be used in this way, we should first allow ourselves to view transitions between states as resulting from the (possibly counterfactual) manifestations of dispositions. The temporal dimension can then be understood as modeling the *higher order* counterfactual manifestation of dispositions, as explored only informally in [Borghini and Williams, 2008].

We mention that a related development that also argues for the metaphysical importance of branching time possibility is presented in [Müller, 2012]. Here, however, the suggestion is made that branching time possibility is in itself metaphysically basic, in that it gives rise to the *real* notion of metaphysical possibility, which, albeit not as wide as that usually considered, is still wide enough to cover the interesting cases, including those that deserve primary attention in metaphysics.

We will now present a case-study which take this point of view further, as an illustration of the potential inherent in this line of research. We will show, in particular, how alternating-time temporal logic (ATL) can be viewed as a theory of dispositional possibility. This will also serve to highlight how the application

[5] Multi-modal logics is a rich topic which is being studied from many different angles and it attracts much technical interest, see [Kurucz et al., 2003].

of branching time systems to study dispositional possibility has the potential to shed light on a number of different, but related, questions, such as the relationship between free will and determinism [List, 2013, Strawson, 1962], the workings of higher order dispositions [Borghini and Williams, 2008], the applicability of notions involving moral responsibility [Frankfurt, 1969, Broersen, 2011a], the nature of necessity and the question of whether or not dispositional possibility is a distinct form of modality [Mumford and Anjum, 2011, Fine, 1994, Fine, 1995], and the distinction between knowing that it is possible to do something, and actually knowing *how* to do it [Jamroga and van der Hoek, 2004, Jamroga and Ågotnes, 2006].

4.1 ATL as a logic of dispositional possibility

In this section we sketch a technical approach to dispositions using ATL, highlighting how the semantics components of this logic can be given a dispositional reading. We also present some ideas for technical developments that suggest themselves on such a reading.

The semantics of ATL is typically given in terms of *concurrent game structures* (CGS's), which can be defined as follows.

Definition 4.1 *A* CGS *is a tuple* $S = \langle \Sigma, Q, \Pi, ((\mathbb{A}_{q,i})_{q \in Q, i \in \Sigma}), \pi, \delta \rangle$ *where:*

- Σ *and* Π *are sets of atoms (usually thought of as agents and propositions respectively, but we will broaden the interpretation of* Σ *in this paper and view it as a collection of arbitrary* object names*).*
- Q *is a non-empty set of states.*
- $\pi : Q \to 2^{\Pi}$ *maps each state to the set of atomic propositions that are true at that state.*
- *For all* $q \in Q, i \in \Sigma$, $\mathbb{A}_{q,i}$ *is a set of atoms associated with* i *at* q*. It is typically thought of as the set of* actions *available for agent* i *at state* q*, but we will broaden the interpretation and view* $\mathbb{A}_{q,i}$ *as a set of* dispositions *for the (possible) object* i*.*
- δ *is a* transition function*. For each* $q \in Q$ *and any tuple* $s \in \prod_{i \in \Sigma} \mathbb{A}_{q,i}$ *(associating an element of* $\mathbb{A}_{q,i}$ *to every* $i \in \Sigma$*) it returns a new state* $q' = \delta(q, s) \in Q$*, referred to as a* successor *of* q*.*

To reason about structures of this kind, a multi-modal language is typically used, which allows us to speak about temporal properties and their interactions with the causal properties of the system, encoded by the transition function, and dependent on how elements of Σ attach themselves to elements of $((\mathbb{A}_{q,i})_{q \in Q, i \in \Sigma})$. More concretely, the transition function depends on what actions agents choose to perform, or, on our reading, on the combinations of dispositions of objects that get triggered in such a way that they manifest. This may or may not be determined – the theory does not compel us to adopt a

particular view on determinism – but from the point of view of what is possible, a state may not admit a unique collection of dispositions that will necessarily manifest. As a result, the corresponding notion of logical time – corresponding here to metaphysical possibility – is in general branching, even if actual time may well not be.

Crucially, a new state is not merely some primitive object, like a possible world or a possible future point in time, but rather a concrete state of affairs which could potentially be brought about causally, as a result of a process that can be traced back to the present state and analyzed as such.[6]. This allows us to express new and interesting properties of possibility that we can not talk about using a standard Kripkean semantics.

In particular, the language of simple ATL, which we will use in this paper, is $\mathcal{L}_{\text{ATL}}$, which we can define by the following grammar:

$$\phi ::= \ p \mid \neg\phi \mid \phi \vee \phi \mid \langle\langle C \rangle\rangle \bigcirc \phi \mid \langle\langle C \rangle\rangle \Box \phi \mid \langle\langle C \rangle\rangle \phi \mathcal{U} \phi$$

where p is a propositional symbol, and $C \subseteq \Sigma$ is a subset of objects from Σ. Intuitively, the language is to be understood as follows:

- $\bigcirc$, $\Box$ and $\mathcal{U}$ are standard *temporal* operators known from many temporal logics, and stand for "next state", "some future state" and "until", respectively;
- $\langle\langle C \rangle\rangle$ is an *ability* operator, and its intuitive meaning is that the set of dispositions attached to objects in C can, irrespectively of what happens to other objects in Σ, *cause* the truth of some formula ϕ which occurs under the scope of one of the temporal operators (i.e., C can cause ϕ to be true eventually, in the next state, or until some other formula ψ becomes true).

For this position paper, we omit a formal definition of truth of $\mathcal{L}_{\text{ATL}}$, but we note that the language of $\mathcal{L}_{\text{ATL}}$ allows us to express interesting properties of causal relations and how they interact with temporal modalities. As we have mentioned, the standard understanding of the parameters in ATL have been that Σ is a collection of agents and that $\mathbb{A}_{q,i}$'s are sets of actions. Then the operator $\langle\langle C \rangle\rangle$ for $C \subseteq \Sigma$ can be understood as expressing the *strategic ability* of the *coalition* C. Under this understanding, the logic of ATL has received much attention, especially from the artificial intelligence community, and in [Goranko and van Drimmelen, 2006] a sound and complete axiomatization was provided. Moreover, epistemic and normative extensions of the logic have been considered, see e.g., [van der Hoek and Wooldridge, 2003, van der Hoek et al., 2006]. For future work, we suggest that these results should be considered from the point of view of the dispositional theory. The question

[6] We also note that while the language of ATL that is presented here only allow us to talk about the future development of the system, the semantics of CGS's allows us to analyze the present in a similar way, as having been caused by processes in one actual among many possible pasts

of whether the axioms for ATL are appropriate also for a theory of dispositional possibility is particularly interesting and should be considered first.

Below we give some examples of ATL-formulas, and their corresponding dispositional reading.

- $\langle\langle C\rangle\rangle \bigcirc \phi$ – there is a (partial) disposition supported by dispositional properties of the objects in C such that ϕ is true in any state where it manifests.
- $\langle\langle C\rangle\rangle \Box \phi$ – there is a (partial, higher-order) disposition supported by a sequence of dispositional properties of C such that ϕ remains true wherever it manifests.
- $\langle\langle C\rangle\rangle \phi \mathcal{U} \psi$ – there is a (partial, higher-order) disposition supported by a sequence of dispositional properties of C such that ϕ is true until, eventually, ψ is true.
- $\langle\langle C\rangle\rangle \Diamond \phi ::= \langle\langle C\rangle\rangle \top \mathcal{U} \phi$ – there is a (partial, higher-order) disposition supported by a sequence of dispositional properties of C such that whenever it manifests, we eventually get ϕ.

We can also define possibility that is general, i.e., not arising from any particular object but rather from the totality of objects. Moreover, we can express different senses in which it is possible for an object to act causally on another object, as illustrated below.

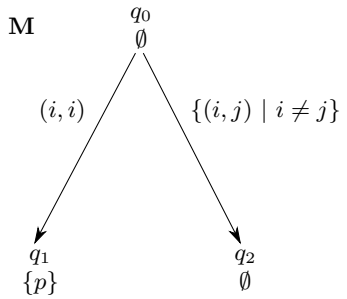
- Metaphysical possibility: $\Diamond \phi$ if, and only if, $\langle\langle \Sigma\rangle\rangle \Diamond \phi$ is true at the actual state q.
- Possible properties of objects can be expressed *without using predicates*:
  - It is actually possible for x to break the glass g: $\langle\langle x\rangle\rangle \Diamond g_{break}$
  - It is possible that it could be actually possible for x to break the glass: $\langle\langle \Sigma\rangle\rangle \Diamond \langle\langle x\rangle\rangle \Diamond g_{break}$

Notice how formalization in terms of ATL highlights the following reasoning task that seems closely associated with the dispositional account: When exactly is it correct to say that a given collection of objects has a possible property? Of course, if one only wishes to speak about possibilities of the world, involving potentially all objects, this problem does not arise. But as soon as one wishes to know more specifically *what* the relevant objects are, the question becomes that of finding a *minimal* collection of objects such that their manifestations suffice to ensure ϕ. In this case one might say more accurately that it is these objects that matter and that genuinely have the property that they render ϕ possible.

For instance, it is not really appropriate to say that it is a property of the glass that it can break, since glasses do not break spontaneously. Rather, it may be the property of some other object x that it can break it, for instance if he is an autonomous agent which moves around in his environment. On the other

hand, it is indeed a property of the glass that it *may* possibly break, which can also be expressed in ATL as $\neg\langle\langle g\rangle\rangle\Box\neg g_{break}$ – there is no disposition such that if it manifests it is impossible for the glass to break.

To further illustrate our perspective we consider an example concerning the relationship between knowledge and ability that has been considered in the ATL literature [Jamroga and van der Hoek, 2004, Jamroga and Ågotnes, 2006]. The scenario is that of a safe and a thief, with the thief lacking knowledge of the code and hence being unable to open the safe. Still, in every possible state he is in some sense able to open it, by simply using the correct code. The problem is that he considers more than one such state possible (since he does not know the code), hence does not know what to do. In the standard way of modeling this, the safe is not modeled as an object in the same way as the agent is. Rather, the different codes the safe might have correspond to different actual states of the world that the thief might be in. On a dispositional reading, on the other hand, there is no reason not to model also the safe as an object – after all, both exist in the actual world. This leads us to view the problem of the thief and the safe as a kind of coordination problem, as sketched below (where s is the safe and t is the thief).



The figure illustrates a model expressing that if the agent chooses to go for the code i and i is indeed the code – among all the possible codes that the safe could have – the door opens, and only then. Below follows some true claims about this model, where p is the propositional atom expressing that the door of the safe is open.

$M, q_0 \models \neg\langle\langle t\rangle\rangle \bigcirc p$ it is impossible for the thief to ensure that the safe opens...
$M, q_0 \models \neg\langle\langle s\rangle\rangle \bigcirc \neg p$ ...but he might get lucky...
$M, q_0 \models \langle\langle s,t\rangle\rangle \bigcirc p$ ...hence opening the safe is a possibility in q_0

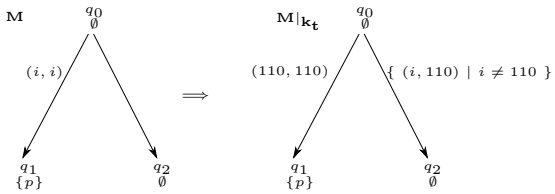
In other words, we can verify formally that unless the thief has more power or knowledge, opening the safe is only one among many possibilities, but also that it depends only on the dispositional properties of the thief and the safe (since any other objects that might be present in q_0 are irrelevant as witnesses to the possibility of p)

Following up on this perspective we may now ask if it suggests new ways of modeling knowledge of dispositional possibility structures. One idea that

suggests itself as soon as we model all objects explicitly is to view knowledge as restriction on what an agent considers possible for other agents and objects, specifically regarding their dispositions and which of them might come to manifest. We leave formal exploration of this idea for future work but make two observations. First, we notice that it renders the signature of knowledge similar to the signature of what is known in the ATL literature as a *normative system* or a social law, see e.g. [van der Hoek et al., 2006]. This serves to highlight the conceptual connection between knowledge and power, and it also suggests directions for technical research on this approach.

Second, we note that the main formal challenge for such an approach to knowledge seems to be to model agents' knowledge of the knowledge of other agents. This can not be done by a straightforward restriction of the models using established techniques from normative systems, but requires instead a more flexible approach which allows us to restrict the models according to the agents' knowledge dynamically and non-monotonically. This, in turn, suggest possible fruitful exchange of ideas with research on the interaction and conflict between different norms, as well as regarding the revision and online design of norms.

Despite the early stages of this research we can give an idea of how such an approach to knowledge will look like, returning to the case of the safe and the thief. Below, we depict a version of the model where it is assumed that the thief knows that the code is 110. Then, from his point of view, the model is restricted such that all other possibilities have to be disregarded, leaving us with the model on the right (k_t denotes the knowledge of the thief, used to update the model).



Here the thief has knowledge, and it is de re knowledge, he knows *how* to open the safe since in the restricted model, he *can*. Turning this idea into a logic of knowledge of ability and dispositional possibility more generally seems like an interesting direction for future work that can also offer a new perspective on the problems associated with modeling de re knowledge of ability in ATL.

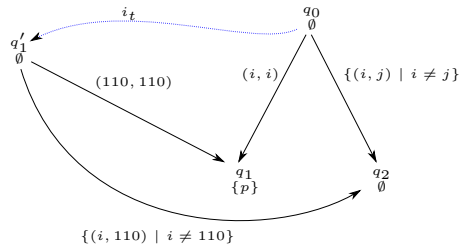
Before we conclude we would like to return to the first form of cognitive activity that was mentioned in the paper, namely contemplation. What, in this technical context, could give substance to a wider form of reasoning about the possibilities that are more like distant images, vague feelings and imprecise goals? Are they relevant at all?

In fact, it seems to us that the epistemic modality of contemplation does have a role to play in this context. Moreover, it seems that it can be formalized using epistemic relations between worlds that are not structured directly by the properties of objects that must be known or otherwise controlled by the

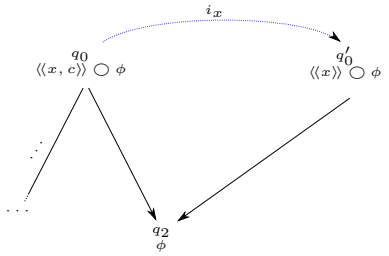
agent, but rather taken to represent his ability to *abstract* from the limitations of the actual and the physical to consider the wider context of possibility within which he is situated. Again we will only sketch the idea, and we will do so by considering the thief again, but in this case such that he does not know what the code is, only contemplating on the possibility that he might come to know, and what it would take for him to arrive in such a more advantageous state.

This can be modeled using a primitive relation of contemplation which directly connect states that might not obviously be connected by any sequence of manifestations. Hence it would allow contemplation about possibilities that are more distant, and for which the main question under the dispositional account would be: how could they come to be manifested? This perspective allows to draw links between the dispositional theory and another aspect of current work on multi-agent formalisms, known as the problem of *synthesis*, finding a concrete strategy for reaching a specific goal.

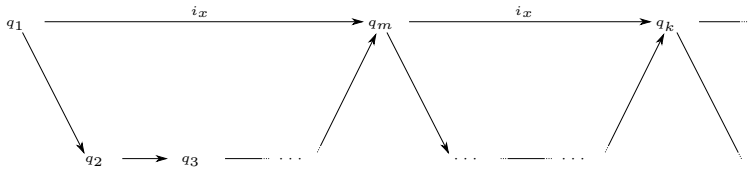
For strategic logics, in particular, this is often flagged as a crucial problem, arising from asking how agents should act in order to bring about a given desired outcome, which may or may not already have been established as a possibility. But mere possibility is of little use in practice unless one knows also *how* it might come about, and this observation applies much more generally, not only to agents' actions but also to other kinds of dispositions. What sequence of dispositions need to be triggered in order for a given possibility to become a reality? Below we depict an example of contemplation where the thief who does not have knowledge about the code *imagines* that the code is 110, modeled by the agent-indexed relation i_t, and is hence allowed to conclude that if this was the case he could open the safe.



More generally, the pattern we have here is an instance of a situation that occurs whenever an agent recognizes that something is possible only if he has control over (or cooperates with) some additional object c and wonders if he may be able to dispense of the need for relying on c to bring about his goal. Then we would model this, quite generally, using the pattern below.



In addition to being a means for formally modeling agents who wonder about what is possible in the dispositional network of the actual world, this approach will also allow us to consider patterns that encode heuristics for establishing how to bring about a given possibility. It could be, for instance, that a given set of sub-goals and preliminary states of affairs might be helpful to consider explicitly, even if they still represent helpful abstraction from the underlying dispositional properties of the objects involved. This would then give rise to patterns such as the one below.



In future work we would like to consider such patterns in more formal detail, in order to further study the interaction between different forms of contemplation and the actual possibilities of physical objects and interacting agents. Moreover, we think that the brief sketch given in this section is enough to suggest that this work can be carried out looking also to work that is currently being carried out on a number of specific issues that arises in the study of multi-agent systems.

5 Conclusion

The primary aim of this paper has been to make a methodological point: since many important questions regarding formal models of social reality involve the relationship between agency and metaphysical possibility, we think more work should be devoted to studying them in this light. We began by giving an introduction to metaphysical modality, arguing that there is nothing mystical about it and that it should be considered. It denotes a form of possibility that is wide enough to cover cases that cannot be completely explained or understood in terms of processes of which we currently possess exact knowledge or predictive power. As such, metaphysical possibility is perhaps the most important kind; it is contemplation about what we don't necessarily understand or are able to describe, and it is exactly this kind of contemplation that can lead to new discoveries. It may involve far fetched stretches of the imagination, but at least

on the actualistic account, to which we adhere, contemplation on metaphysical possibility also comes with a commitment to search for foundations in the actual world.

We went on to describe the dispositional account of possibility, a metaphysical theory which appears to adopt just such a measured stance on what possibility is, and what it is good for. Instead of starting with the possible, this theory starts with the actual, and it posits that anything that is possible, even in the metaphysical sense, can in principle be traced back to actuality by identifying the sequences of dispositions that would have to manifest to bring it about. We argued that this strongly suggests formal representation using branching time formalisms, pointing out that these are much studied in logics for artificial intelligence. We went on to provide a more elaborate technical case-study, giving a dispositional reading of the strategic multi-agent logic ATL. We also took the opportunity to suggest some possible benefits that could arise from doing this, and presented several ideas for further technical work.

The continuous exchange of ideas between different fields of research has become one of the defining features of the community of researches who employ formal tools to study social reality. Hopefully, we made a good case in this paper for the claim that the distinct notion of metaphysical possibility should not be overlooked in this regard. In particular, we think the recently introduced dispositional theory serves to illustrate this point nicely, showing that metaphysics should be welcomed to the fold.

References

[Ågotnes et al., 2009] Ågotnes, T., van der Hoek, W., and Wooldridge, M. (2009). Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30.

[Alur et al., 2002] Alur, R., Henzinger, T., and Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5):672–713.

[Belnap and Perloff, 1988] Belnap, N. and Perloff, M. (1988). Seeing to it that: A canonical form for agentives. *Theoria*, 54(3):175–199.

[Borghini and Williams, 2008] Borghini, A. and Williams, N. E. (2008). A dispositional theory of possibility. *Dialectica*, 62(1):21–8211.

[Bricker, 2008] Bricker, P. (2008). Concrete possible worlds. In *Contemporary Debates in Metaphysics*. Blackwell Pub.

[Broersen, 2011a] Broersen, J. (2011a). Deontic epistemic stit logic distinguishing modes of mens rea. *J. Applied Logic*, 9(2):137–152.

[Broersen, 2011b] Broersen, J. (2011b). Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic*, 40(4):499–530.

[Epstude and Roese, 2008] Epstude, K. and Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):68–92.

[Fine, 1994] Fine, K. (1994). Essence and modality. *Philosophical Perspectives*, 8:1–16.

[Fine, 1995] Fine, K. (1995). The logic of essence. *Journal of Philosophical Logic*, 24(3):241–273.

[Frankfurt, 1969] Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(3):829–39.

[Goranko and van Drimmelen, 2006] Goranko, V. and van Drimmelen, G. (2006). Complete axiomatization and decidability of alternating-time temporal logic. *Theor. Comput. Sci.*, 353(1-3):93–117.

[Horty and Belnap, 1995] Horty, J. F. and Belnap, N. (1995). The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644.

[Jacobs, 2010] Jacobs, J. D. (2010). A powers theory of modality: or, how I learned to stop worrying and reject possible worlds. *Philosophical Studies*, 151:227–248.

[Jamroga and Ågotnes, 2006] Jamroga, W. and Ågotnes, T. (2006). What agents can achieve under incomplete information. In Stone, P. and Weiss, G., editors, *Proc. of the Fifth Intern. Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 232–234. ACM Press.

[Jamroga and van der Hoek, 2004] Jamroga, W. and van der Hoek, W. (2004). Agents that know how to play. *Fundamenta Informaticae*, 63:185–219.

[Joyce, 1999] Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

[Kripke, 1981] Kripke, S. (1981). *Naming and Necessity*. Blackwell Publishing.

[Kripke, 2005] Kripke, S. (2005). Identity and necessity. In Loux, M. J., editor, *Metaphysics - Contemporary Readings*. Routledge.

[Kurucz et al., 2003] Kurucz, A., Wolter, F., Zakharyaschev, M., and Gabbay, D. M. (2003). *Many-Dimensional Modal Logics: Theory and Applications*, volume 148 of *Studies in Logic and the Foundations of Mathematics*. Elsevier.

[Leibniz, 1998] Leibniz, G. (1998). *Theodicy*. Open Court. First published in 1709.

[Lewis, 1971] Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68(7).

[Lewis, 1986] Lewis, D. (1986). *On the plurality of worlds*. Oxford University Press.

[List, 2013] List, C. (2013). Free will, determinism, and the possibility of doing otherwise. *Noûs*, 47(2).

[Müller, 2012] Müller, T. (2012). Branching in the landscape of possibilities. *Synthese*, 188:41–65.

[Mumford and Anjum, 2011] Mumford, S. and Anjum, R. L. (2011). Dispositional modality. In *Lebenswelt und Wissenschaft, Deutsches Jahrbuch Philosophie 2*. Meiner Verlag.

[Parikh, 2001] Parikh, R. (2001). Social software. *Synthese*, 132:200–2.

[Rosen, 1990] Rosen, G. (1990). Modal fictionalism. *Mind*, 99(395):327–354.

[Rosen, 1995] Rosen, G. (1995). Modal fictionalism fixed. *Analysis*, 55(2):67–73.

[Strawson, 1962] Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48:1–25.

[van Benthem, 2011] van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[van der Hoek et al., 2006] van der Hoek, W., Roberts, M., and Wooldridge, M. (2006). Social laws in alternating time: effectiveness, feasibility, and synthesis. *Synthese*, 156(1):1–19.

[van der Hoek and Wooldridge, 2003] van der Hoek, W. and Wooldridge, M. (2003). Cooperation, knowledge and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75:125–157.

[Vetter, 2010] Vetter, B. (2010). *Potentiality and possiblity*. University of Oxford. PhD thesis.

[Vetter, 2011] Vetter, B. (2011). Recent work: Modality without possible worlds. *Analysis*, 71(4):742–754.

[Vetter, 2013] Vetter, B. (2013). 'can' without possible worlds: Semantics for anti-humeans. *Philosophers' Imprint*, 13(16).

[Wooldridge, 2009] Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. John Wiley & Sons, Inc., 2 edition.