The IfCoLog Journal of Logics and their Applications

Volume 1 • Issue 2 • November 2014

Contents

ISBN 978-1-84890-163-6
90000
9 781848 901636

Volume 1 ● Issue 2 ● November 2014

The IfColog Journal of Logics and their applications

The IfCoLog Journal of Logics and their Applications

Volume 1 ● Issue 2 ● November 2014

9 781848 901636

7.44 x 9.69
246 mm x 189 mm

.309
7.848mm

7.44 x 9.69
246 mm x 189 mm

Content Type: Black & White
Paper Type: White
Page Count: 144
File Type: InDesign
Request ID: CSS1283954

Perfect Bound Cover Template

Lightning Source

Document Size: 19" x 12"
305 x 483mm

IFCoLog Journal of Logics and their Applications

Volume 1, Number 2

November 2014

Disclaimer
Statements of fact and opinion in the articles in IfCoLog Journal of Logics and their Applications are those of the respective authors and contributors and not of the IfCoLog Journal of Logics and their Applications or of College Publications. Neither College Publications nor the IfCoLog Journal of Logics and their Applications make any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

ISBN 978-1-84890-163-6
ISSN (E) 2055-3714
ISSN (P) 2055 3706

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Printed by Lightning Source, Milton Keynes, UK

Editorial Board

Scope and Submissions

This journal considers submission in all areas of pure and applied logic, including:

pure logical systems
proof theory
constructive logic
categorical logic
modal and temporal logic
model theory
recursion theory
type theory
nominal theory
nonclassical logics
nonmonotonic logic
numerical and uncertainty reasoning
logic and AI
foundations of logic programming
belief revision
systems of knowledge and belief
logics and semantics of programming
specification and verification
agent theory
databases

dynamic logic
quantum logic
algebraic logic
logic and cognition
probabilistic logic
logic and networks
neuro-logical systems
complexity
argumentation theory
logic and computation
logic and language
logic engineering
knowledge-based systems
automated reasoning
knowledge representation
logic in hardware and VLSI
natural language
concurrent computation
planning

This journal will also consider papers on the application of logic in other subject areas: philosophy, cognitive science, physics etc. provided they have some formal content.

Submissions should be sent to Jane Spurr (jane.spurr@kcl.ac.uk) as a pdf file, preferably compiled in LaTeX using the IFCoLog class file.

Contents

ARTICLES

Deontic Logic and Preference Change

Johan van Benthem
University of Amsterdam and Stanford University

Fenrong Liu
Tsinghua University, Beijing, China

Abstract

The normative realm involves deontic notions such as obligation or permission, as well as information about relevant actions and states of the world. This mixture is not static, given once and for all. Both information and normative evaluation available to agents are subject to changes with various triggers, such as learning new facts or accepting new laws. This paper explores models for this setting in terms of dynamic logics for information-driven agency. Our paradigm will be dynamic-epistemic logics for knowledge and belief, and their current extensions to the statics and dynamics of agents' preferences. Here the link with deontics is that moral reasoning may be viewed as involving preferences of the acting agent as well as moral authorities such as lawgivers, one's conscience, or yet others. In doing so we discuss a large number of themes: primitive 'betterness' order versus reason-based preferences (employing a model of 'priority graphs'), the entanglement of preference and informational attitudes such as belief, interactive social agents, and scenarios with long-term patterns emerging over time. Specific deontic issues considered include paradoxes of deontic reasoning, acts of changing obligations, and changing norm systems. We conclude with some further directions, as well as a series of pointers to related work, including different paradigms for looking at these same phenomena.

Keywords: Deontic Logic, Preference Change, Epistemic Logic, Public Announcement Logic.

This paper is an updated and revised version of a draft chapter for the *Handbook of Deontic Logic*, that was written originally in 2009. Given the recent increasing interest in our central themes of reason-based preference and preference dynamics, we are publishing the present version in the IF-COLOG Journal at the kind suggestion of Dov Gabbay. We are grateful to Guillaume Aucher and Davide Grossi for many useful comments and pointers, the majority of which will feed into the final chapter version when the Handbook appears.

1 Agency, information, and preference

Agents pursue goals in this world, acting within constraints in terms of their information about what is true, as well as norms about what is right. The former dimension typically involves acts of inference, observation, as well as communication and other forms of social interaction. The latter dimension involves evaluation of situations and actions, 'coloring' the agents' view of the world, and driving their desires, decisions, and actions in it. A purely informational agent may be rational in the sense of clever reasoning, but a *reasonable* agent is one whose actions are in harmony with what she wants. The two dimensions are intimately related. For instance, what we want is influenced by what we believe to be true as well as what we prefer, and normally also, we only seek information to further goals that we desire.

This balance of information and evaluation is not achieved once and for all. Agents must constantly cope with new information, either because they learn more about the current situation, or because the world has changed. But equally well, agents constantly undergo changes in evaluation, sometimes by intrinsic changes of heart, but most often through events with normative impact, such as accepting a command from an authority. These two forms of dynamics, too, are often entangled: for instance, learning more about the facts can change my evaluation of a situation.

A third major aspect of agency is its social interactive character. Even pure information flow is often driven by an epistemic gradient: the fact that different agents know different things leads us to communicate, whether in cooperative inquiry or adversarial argumentation, perhaps until a state of equilibrium is reached such as common knowledge or common belief. But also more complex forms of interaction occur, such as merging beliefs, where differences in informational authority may play a crucial role. Again, very similar phenomena play on the normative side. Norms, commitments and duties usually involve other agents, both as their source and as their target, and whole institutions and societies are constructed in terms of social choice, shared norms and rules of behavior.

In this chapter, we will discuss how current dynamic-epistemic logics can model the above phenomena, both informational and preferential, and we will show what results when this perspective is taken to normative reasoning and deontic logic. Our treatment will be brief, and for a much more elaborate sample of this style of thinking about the normative realm, we refer to [17]. In pursuing this specific line, we are not denying the existence of other valid approaches to deontic dynamics, and we will provide a number of references to other relevant literature.

2 Dynamic logics of knowledge and belief change

Before analyzing preference or related deontic notions, we first develop the basic methodology of this paper for the purely informational case, where the first 'dynamic-epistemic logics' arose in the study of information change.

2.1 Epistemic logic and semantic information

Dynamic logics of agency need an account of underlying of static states that can be modified by suitable triggers: actions or events. Such states usually come from existing systems in philosophical or computational logic whose models can serve as static snapshots of the dynamic process. In this paper, we start with a traditional modal base system of epistemic logic, referring to the standard literature for details (cf. [45] and [30]).

Definition 1. *Let a set of propositional variables Φ be given, as well as a set of agents A. The epistemic language is defined by the syntax rule*

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \qquad \textit{where } p \in \Phi,\ a \in A.$$

Remark: Single agents, interacting agents, and groups. For convenience, we will focus on single agents in this paper, although this still allows us to describe interacting individual agents where needed through iterations of modalities. Epistemically important notions with groups themselves as agents, such as 'common knowledge' or 'distributed knowledge', are deferred to our discussion at the end.

Semantic models for the epistemic language encode agents' 'information ranges' in the form of equivalence classes of binary uncertainty relations for each agent.[1] These support a standard compositional truth definition.

Definition 2. *An* epistemic model *is a tuple $\mathfrak{M} = (W, \{\sim_a\}_{a\in A}, V)$ with W a set of epistemically possible states (or 'worlds'), $\sim_a$ an equivalence relation on W, and V a valuation function from Φ to subsets of W.*

Definition 3. *For an epistemic model $\mathfrak{M} = (W, \{\sim_a \mid a \in A\}, V)$ and any world $s \in S$, we define $\mathfrak{M}, s \models \varphi$ (epistemic formula φ* is true in $\mathfrak{M}$ at *s) by induction on the structure of the formula φ:*

1. *$\mathfrak{M}, s \models \top$ always.*

[1] The approach of this paper will also work on more general relations such as pre-orders, but we start with this easily visualizable epistemic case for expository purposes.

2. $\mathfrak{M}, s \models p$ iff $s \in V(p)$.

3. $\mathfrak{M}, s \models \neg\varphi$ iff *not* $\mathfrak{M}, s \models \varphi$.

4. $\mathfrak{M}, s \models \varphi \wedge \psi$ iff $\mathfrak{M}, s \models \varphi$ *and* $\mathfrak{M}, s \models \psi$.

5. $\mathfrak{M}, s \models K_a\varphi$ iff *for all* t *with* $s \sim_a t : \mathfrak{M}, t \models \varphi$.

Using equivalence relations in our models yields the well-known modal system **S5** for each individual knowledge modality, without interaction laws for different agents. Just for concreteness, we record this basic fact here:

Theorem 4. *Basic epistemic logic is axiomatized completely by the axioms and inference rules of the modal system* ***S5*** *for each separate agent.*

Few researchers see our basic modalities and the simple axioms of modal **S5** as expressing genuine properties of 'knowledge' – thus making earlier polemical discussions of epistemic 'omniscience' or 'introspection' expressed by these axioms obsolete. Our interpretation of the above notions is as describing the *semantic information* that agents have available (cf. [13]), being a modest but useful building block in analyzing more complex epistemic and deontic notions. We will allow ourselves the use of the word 'know' occasionally, however: old habits die hard.[2]

Static epistemic logic describes what agents know on the basis of their current semantic information. But information flows, and a richer story must also include dynamics of actions that produce and modify information. We now turn to the simplest case of this dynamics: reliable public announcements or public observations, that shrink the current information range.

2.2 Dynamic logic of public announcement

The pilot for the methodology of this paper is 'public announcement logic' (PAL), a toy system describing a combination of epistemic logic and one dynamic event, namely, *announcement* of new 'hard information' expressed in some proposition φ that is true at the actual world. The corresponding 'update action' $!\varphi$ transforms a current epistemic model $\mathfrak{M}, s$ into its definable submodel $\mathfrak{M}|\varphi, s$ where all worlds

[2] There is a fast-growing literature on more sophisticated logical analyses of genuine knowledge (cf. [76], [27], [123]), which also seems relevant to modeling and reasoning in the deontic realm. However, the main points to be made in this paper are orthogonal to these additional refinements of the logical framework.

that did not satisfy φ have been eliminated. This model update is the basic scenario of obtaining information in the realm of science but also of common sense, by shrinking one's current epistemic range of uncertainty.[3]

To describe this phenomenon, the *language* of PAL has two levels, using both formulas for propositions and action expressions for announcements:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [A]\varphi$$
$$A := !\varphi$$

The new dynamic formula $[\varphi]\psi$ says that "after updating with the true proposition φ, formula ψ holds":

$$\mathfrak{M}, s \models [!\varphi]\psi \text{ iff if } \mathfrak{M}, s \models \varphi, \text{ then } \mathfrak{M}|\varphi, s \models \psi.$$

This language can make characteristic assertions about knowledge change such as $[!\varphi]K_a\psi$, which states what agent a will know after having received the hard information that φ. In particular, the knowledge change before and after an update can be captured by so-called *recursion axioms*, a sort of recursion equations for the 'dynamical system' of PAL, relating new knowledge to knowledge that agents had before. Here is the complete logical system for information flow under public announcement (two original sources are [53], [111]):

Theorem 5. *PAL is axiomatized completely by the usual laws of the static epistemic base logic plus the following recursion axioms:*

1. $[!\varphi]q \leftrightarrow (\varphi \rightarrow q)$ *for atomic facts q*
2. $[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$
3. $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
4. $[!\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a[!\varphi]\psi)$

These elegant principles analyze reasoning about epistemic effects of receiving hard information, through observation, communication, or other reliable means. In particular, the knowledge law reduces knowledge after new information to 'conditional knowledge' that the agent had before, but in a subtle recursive manner. This prudence of design for PAL is necessary since the process of information update can change truth values of epistemic assertions. Perhaps, initially, I did not know that p, but after the event $!p$, I do.

[3] The name 'public announcement logic' may be unfortunate, since the logic describes updates with hard information from whatever source, but no consensus has emerged yet on a rebaptism.

There are several noteworthy features to this approach. We already stressed the recursive nature of reducing new knowledge to pre-existing knowledge, a feature that is typical of dynamical systems. Also, the precise way in which this happens involves breaking down the 'postconditions' behind the dynamic modalities $[!\varphi]$ compositionally on the basis of their shape.

Next, as things stand here, repeating these steps, the stated features drive a 'reduction process' taking every formula of our dynamic-epistemic language eventually to an equivalent formula inside the static epistemic language. In terms of semantics and expressive power, this means that a current static model 'pre-encodes' all information about what might happen when agents communicate what they know. In terms of the logic, the reduction procedure means that PAL is *axiomatizable* and *decidable*, since it inherits these features from the epistemic base logic.

However, it is also important to note that the latter sweeping dynamics-to-statics reduction is not an inevitable feature of dynamic-epistemic analysis. In recent versions of the semantics for PAL, the available sequences of information updates may be constrained by global *protocols* that regulate available events in the current process of inquiry. In that case, no reduction is possible to the base logic, and the dynamic logic, though still employing recursion equations, while also remaining axiomatizable and decidable, comes to encode a genuine new kind of 'procedural information' (cf. [14]). Protocols also make sense for deontic purposes, because of the procedural character of much normative behavior, and we will briefly return to this perspective at the end of this chapter.

In what follows, PAL will serve as a pilot example for many other complex cases, for example, changes in beliefs, preferences, and obligations. In each case, the 'triggering events' can be different: for instance, beliefs can change by signals of different force: hard or more 'soft', and obligations can change through actions of commanding by a normative authority. In many cases, the domain of the model does not change, but rather its *ordering pattern.*[4] However, the general recursive methodology of PAL will remain in force, though in each case, with new twists.

2.3 From knowledge to belief and soft information

Knowledge rests on hard information, but most of the information that we have and act on is soft, giving rise to *beliefs*, that are not always true, and that can be revised when shown inadequate. One can think of learning from error as the more creative ability, beyond mere recording of reliable information in the agent's environment.

[4] One example of this approach, even in the epistemic realm, are 'link cutting' versions of updating after announcement: cf. [90], [126], [26], that will be used later on in scenarios where we may want to return to worlds considered earlier in the process.

Again we need to start with a convenient static base for our investigation. One powerful model for soft information and belief reflects the intuition that we believe those things that hold in the *most plausible* worlds in our epistemic range. I believe that this train will take me home on time, even though I do not know that it will not suddenly fly away from the tracks. But the worlds where it stays on track are more plausible than those where it flies off, and among the latter, those where it arrives on time are more plausible than those where it does not.

The long history for this way of modeling belief includes non-monotonic logic in artificial intelligence ([124], [32], [86], [50], [51]), the semantics of natural language (cf. [139]), as well as the philosophical literature on epistemology, and logics of games (cf. [129], [10]).

The common intuition of relative plausibility leads to the following semantics:

Definition 6. *An* epistemic-doxastic model $\mathfrak{M} = (W, \{\sim_a\}_{a\in A}, \{\leq_a\}_{a\in A}, V)$ *consists of an epistemic model* $(W, \{\sim_a\}_{a\in A}, V)$ *as before, while the* $\leq_a$ *are binary comparative plausibility pre-orders for agents between worlds.*

Intuitively, these comparison orders might well be *ternary* $\leq_{a,s} xy$ saying that, in world s, agent a considers world x at least as plausible as y.[5] For convenience in this chapter, however, our semantics assumes that plausibility orderings are the same for epistemically indistinguishable worlds: that is, agents know their plausibility judgements. Assuming that plausibility is a pre-order, i.e., reflexive and transitive, but not necessarily connected, leaves room for the existence of genuinely incomparable worlds – but much of what we say in this chapter also holds for the special case of *connected* pre-orders where any two worlds are comparable.[6] As with epistemic models, our logical analysis works largely independently from specific design decisions about the ordering, important though they may be in specific applications.

One can interpret many logical languages in these comparative order structures. In what follows, we work with modal formalisms for the usual reasons of perspicuous formulation and low complexity (cf. [29]).

First of all, there is *absolute belief* as truth in all most plausible worlds:

$\mathfrak{M}, s \models B_a\varphi$ iff $\mathfrak{M}, t \models \varphi$ *for all those worlds* $t \sim_a s$ *that are maximal in the order* $\leq_a xy$ *in the* $\sim_a$*-equivalence class of* s.

[5] In particular, ternary world-dependent plausibility relations are found in the semantics of conditional logic: cf. [89], [127], models for games: cf. [130], [13], as well as in recent logical analyses of major paradigms in epistemology: [76].

[6] Connected orders are equivalent to the 'sphere models' of conditional logic or belief revision theory (cf. [62], [120]) – but in these areas, too, a generalization to pre-orders has been proposed: cf. [35], [124], [138].

But the more general notion in our models is that of a *conditional belief*:

> $\mathfrak{M}, s \models B_a^{\psi}\varphi$ iff $\mathfrak{M}, t \models \varphi$ *for all those worlds* $t \sim_a s$ *that are maximal for* $\leq_a xy$ *in the set* $\{u \mid s \sim_a u$ *and* $\mathfrak{M}, u \models \psi\}$.[7]

Conditional beliefs generalize absolute beliefs, which are now definable as $B_a^{\top}\varphi$. They *pre-encode* absolute beliefs that we will have *if* we learn certain things. Indeed, the above semantics for $B_a^{\psi}\varphi$ is formally similar to that for conditional assertions $\psi \Rightarrow \varphi$. This allows us to use known results from [35], [138]:

Theorem 7. *The logic of* $B_a^{\psi}\varphi$ *is axiomatized by standard propositional logic plus the laws of conditional logic over pre-orders.*

Deductively stronger modal logics also exist in this area, such as the popular system **KD45** for absolute belief. The structural content of their additional axioms can be determined through standard modal frame correspondence techniques (see [29], [23]).

Digression: Further relevant attitudes. Modeling agency with just the notions of knowledge and belief is mainly a tradition inherited from the literature. In a serious study of agency the question needs to be raised afresh what is our natural repertoire of attitudes triggered by information. As one interesting example, the following operator has emerged recently, in between knowledge and belief qua strength. Intuitively, 'safe belief' is belief that agents have which cannot be falsified by receiving true new information.[8] Over epistemic plausibility models $\mathfrak{M}$, its force is as follows:

Definition 8. *The modality of* safe belief $B_a^{+}\varphi$ *is interpreted as follows:*

> $\mathfrak{M}, s \models B_a^{+}\varphi$ iff *for all worlds* $t \sim_a s$*: if* $s \leq_a t$*, then* $\mathfrak{M}, t \models \varphi$.

Thus, the formula φ is to be true in all accessible worlds that are at least as plausible as the current one. This includes the most plausible worlds, but it need not include all epistemically accessible worlds, since the latter may include some less plausible worlds than the current one. The logic for safe belief is just **S4**, since it is in fact the simplest modality over the plausibility order.

[7] These intuitive maximality formulations must be modified in models allowing infinite sequences in the plausibility ordering. Trivialization can then be avoided as follows (cf. the exposition of plausibility semantics in [54]): $\mathfrak{M}, s \models O^{\psi}\varphi$ iff $\forall t \sim s : \exists u :$ ($t \preceq u$ *and* $\mathfrak{M}, u \models \psi$ *and* $\forall v \sim s$: (*if* $u \preceq v$ *and* $\mathfrak{M}, v \models \psi$, *then* $\mathfrak{M}, v \models \varphi$)).

[8] This notion has been proposed independently in AI [125], philosophy [131], learning theory, and game theory [8], [11].

A notion like this has the conceptual advantage of making us see that agents can have more responses to information than just knowledge and belief.[9] But there is also the technical advantage that the simple modality of safe belief can define more complex notions such as conditional belief (see [85], [33], [13]) which can lead to simplifications of logics for agency.

2.4 Dynamic logics of belief change

Having set up the basic attitudes, we now want to deal with explicit acts or events that update not just knowledge, but also agents' beliefs.[10]

Hard information The first obvious triggering event are the earlier public announcements of new hard information. Their complete logic of belief change can be developed in analogy with the earlier dynamic epistemic logic PAL, again via world elimination. Its key recursion axiom for new beliefs uses conditional beliefs:

Fact 9. *The following formula is valid in our semantics:*

$$[!\varphi]B_a\psi \leftrightarrow (\varphi \to B_a^{\varphi}[!\varphi]\psi)$$

To keep the complete dynamic language in harmony, we then also need a recursion axiom for the conditional beliefs that are essential here:

Theorem 10. *The dynamic logic of conditional belief under public announcements is axiomatized completely by*

(a) any complete static logic for the model class chosen,

(b) the PAL recursion axioms for atomic facts and Boolean operations,

(c) the following recursion axiom for conditional beliefs:

$$[!\varphi]B_a^{\chi}\psi \leftrightarrow (\varphi \to B_a^{\varphi\wedge[!\varphi]\chi}[!\varphi]\psi)$$

This analysis also extends to safe belief, with this recursion law:

Fact 11. *The following PAL-style axiom holds for safe belief:*

$$[!\varphi]B_a^{+}\psi \leftrightarrow (\varphi \to B_a^{+}(\varphi \to [!\varphi]\psi)).$$

[9] Other relevant notions include the 'strong belief' of [131], [10].

[10] For a much more extensive up-to-date treatment of logic-based belief revision, cf. the chapter [28] in the forthcoming *Handbook of Logics of Knowledge and Belief.*

Using this equivalence, which behaves more like the original central PAL axiom, one can show that safe belief has its intuitively intended feature. Safe belief in factual propositions (i.e., those not containing epistemic or doxastic operators) remains safe belief after updates with hard factual information.[11]

Soft information But belief change also involves more interesting triggers, depending on the quality of the incoming information, or the trust agents place in it. 'Soft information upgrade' does not eliminate worlds as what hard information does, but rather *changes the plausibility order*, promoting or demoting worlds according to their properties. Here is one widely used way in which this can happen: an act of 'radical', or 'lexicographic' upgrade.[12]

Definition 12. *A* radical upgrade $\Uparrow\varphi$ *changes the current plausibility order* $\leq$ *between worlds in* $\mathfrak{M}, s$ *to create a new model* $\mathfrak{M}{\Uparrow}\varphi, s$ *where all* φ*-worlds in* $\mathfrak{M}, s$ *become better than all* $\neg\varphi$*-worlds, while, within those two zones, the old plausibility order* $\leq$ *remains as it was.*

No worlds are eliminated here, it is the ordering pattern that adapts. There is a matching upgrade modality for this in our dynamic language:

$\mathfrak{M}, s \models [\Uparrow\varphi]\psi$ iff $\mathfrak{M}{\Uparrow}\varphi, s \models \psi$.

This supports one more dynamic completeness theorem (cf.[22]).

Theorem 13. *The logic of radical upgrade is axiomatized completely by*

(a) a complete axiom system for conditional belief on the static models,

(b) the following recursion axioms:

$$\begin{aligned}
&[\Uparrow\varphi]q \;\leftrightarrow\; q, \quad \textit{for all atomic proposition letters } q\\
&[\Uparrow\varphi]\neg\psi \;\leftrightarrow\; \neg[\Uparrow\varphi]\psi\\
&[\Uparrow\varphi](\psi \wedge \chi) \;\leftrightarrow\; ([\Uparrow\varphi]\psi \wedge [\Uparrow\varphi]\chi)\\
&[\Uparrow\varphi]B^{\chi}\psi \;\leftrightarrow\; (E(\varphi \wedge [\Uparrow\varphi]\chi) \wedge B^{\varphi\wedge[\Uparrow\varphi]\chi}[\Uparrow\varphi]\psi)\\
&\qquad \vee (\neg E(\varphi \wedge [\Uparrow\varphi]\chi) \wedge B^{[\Uparrow\varphi]\chi}[\Uparrow\varphi]\psi
\end{aligned}$$

[11] Unlike with plain belief, the latter recursion does not involve a move to an irreducible new notion of 'conditional safe belief'. Indeed, given a definition of conditional belief in terms of safe belief, the more complex recursion law in Theorem 10 can be derived.

[12] In this section, we drop epistemic accessibility, and focus on plausibility order only.

Here the operator 'E' is the existential epistemic modality, and we need to add a simple recursion axiom for knowledge, that we forego here.[13]

There are many further policies for changing plausibility order whose dynamic logic can be axiomatized in a similar manner. For instance, 'conservative upgrade' $\uparrow\varphi$ only puts the *most plausible* φ-worlds on top in the new model, leaving the rest in their old positions. For general results on complete logics, see [22], [10] and [12]. In particular, [117] is an excellent source for variety of policies in belief revision theory that is not tied to the specific dynamic logic methodology employed in this paper.

2.5 General dynamic methodology and its applications

We have spent quite some time on the above matters because they represent a general methodology of *model transformation* that works for many further phenomena, including changes in preference, and the even richer deontic scenarios that we will be interested in eventually.

Model transformations of relevance to agency can be much more drastic than what we have seen here, extending the domains of available worlds and modifying their relational structure accordingly. In the dynamic-epistemic logic of general observation DEL, different agents can have different access to the current informational event, as happens in card games, communication with security restrictions, or other social scenarios. This requires generalizing PAL as well as the above logics of belief change, using a mechanism of 'product update' to create more complex new models (cf. [9], [136], [12]).

Appropriately extended update mechanisms have been applied to many further aspects of agency: changes in intentions ([118], [80]), trust ([75]), inference ([137]), questions and inquiry ([19]), as well as complex scenarios in games ([105], [13]) and social information phenomena generally ([121], [7], [66]). Yet, in this paper, we will stick mainly with the much simpler pilot systems presented in the preceding sections.

3 Deontic logic as preference logic

Having set up the machinery for changing informational attitudes, we now turn to our next interest, the realm of normative evaluation for worlds or actions and the matching dynamic deontic logics. Here we will follow a perhaps not uncontroversial track: our treatment of deontic notions and scenarios will be based on *preference* structure and its changes. We believe that this is a conceptually good way of looking at deontic notions, and at the same time, it lends itself very well to treatment by our

[13] As before, it is easy to extend this analysis of soft upgrade to safe belief.

earlier methods, since at an abstract level, doxastic plausibility order and deontic betterness order are very similar. The results that follow in the coming sections are largely from [91], [54], and [94].[14]

Let us say a few more words about the connection between deontic logic and preference, to justify our approach in this paper. Deontic logic is the logical study of normative concepts such as obligation, prohibition, permission and commitment. This area was initiated by von Wright in [140] who introduced the logic of absolute obligation. As a reaction to paradoxes with this notion, conditional obligation was then proposed in [141], [143] and [47]. Good reviews systematizing the area are found in [3], [4].

One often thinks of deontic logic as the study of some accessibility relation from the actual world to the set of 'ideal worlds', but the more sophisticated view ([67], [48] and [81]) has models with a binary comparison relation that we may call 'betterness'.[15] Such more general comparisons make sense, for instance, when talking and reasoning about 'the lesser of two evils', or about 'improvement' of some given situation.

Naturally, this is precisely the ordering semantics that we have already seen for belief, and it would be tedious to indulge in formal definitions at this stage that the reader can easily construct for herself. Our base view would be that of binary *pre-orders* as before, for which we will now use the notation R to signal a change from the earlier plausibility interpretation. As usual, imposing further constraints on the ordering will generate deductively stronger deontic logics.

The binary relation R now interprets $O\varphi$ (absolute obligation) as φ *being true in all best worlds*, much like belief with respect to plausibility. Likewise, we interpret conditional obligation $O^{\psi}\varphi$ like conditional belief: φ holds *in the best* ψ*-worlds*.[16]

For further information on deontic logic, we refer to [4] and various chapters in the forthcoming Handbook [52]. Our emphasis in this paper will be mainly on interfacing with this field.

As we already noted at the start of this paper, deontic ordering shows intuitive analogies with the notion of *preference*. One can think of betterness as reflecting the preferences of a moral authority or law-giver, and in the happy Kantian case

[14] To unclutter notation, here and henceforth, we will mostly suppress agent indices for modal operators and their corresponding relations.

[15] Hansson argued that von Wright-type deontic logic can be naturally interpreted in terms of a preference relation 'is at least as ideal as' among possible worlds – an ordering that we will call 'betterness' in what follows.

[16] There are also more abstract neighborhood versions of this semantics, where the current proposition plays a larger role in terms of binary deontic betterness relations R^{ψ}, where one can set $\mathfrak{M}, s \models O^{\psi}\varphi$ iff for all t in W with $sR^{\psi}t, \mathfrak{M}, t \models \varphi$.

where agents' duties coincide with their inclinations, deontic betterness *is* in fact the agent's own preference. We claim no novelty for this line of thought, which was advocated forcefully as early as [67]. With this twist, we can then avail ourselves of existing studies of preference structure and evaluation dynamics, a line of thinking initiated in [134] and [135], though we now take the dynamic-epistemic road.

By way of background to what follows, we note that preference logic is a vigorous subject with its own history. For many new ideas and results in the area, we refer to [71] and [63]. What we will do next in this paper is survey some recent developments in the study of preference statics and dynamics, emphasizing those that are of relevance to deontic logic, an area where we will return eventually toward the end of this paper.[17]

4 Static preference logic

In the coming sections, we will discuss basic developments in modal preference logic, starting with its statics, and then continuing with the dynamics of preference change. Our treatment follows ideas from [33] and [65], and for the dynamics, we rely on [20] and [26].

4.1 General modal preference logic

Our basic models are like in decision theory or game theory: there is a set of alternatives (worlds, outcomes, objects) ordered by a primitive ordering that we dub 'betterness' to distinguish it from richer notions of preference.[18]

Definition 14. *A* modal betterness model *is a tuple* $\mathfrak{M} = (W, \preceq, V)$ *with* W *a set of worlds or objects,* $\preceq$ *a reflexive and transitive relation over these, and* V *is a valuation assigning truth values to proposition letters at worlds.*[19]

The order relation in these models also induces a strict variant $s \prec t$:

If $s \preceq t$ but not $t \preceq s$, then t is *strictly better* than s.

[17] Preference logic tends to focus on describing the agents' own preferences, rather than those of others, but what we have to say applies equally well to multi-agent settings such as moral scenarios, or games, where different preference orders interact in crucial ways.

[18] To repeat an earlier point, while each agent has her own betterness order, in what follows, merely for technical convenience, we suppress indices wherever we can.

[19] As we said before, we use pre-orders since we want the generality of possibly non-total preferences. Still, total orders, the norm in areas like game theory, provide an interesting specialization for the results in this chapter – but we will only mention it in passing.

Here is a simple modal language that can say a lot about these structures:

Definition 15. *Take any set of propositional variables Φ, with p ranging over Φ. The* modal betterness language *has this inductive syntax rule:*

$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\leq\rangle\varphi \mid \langle<\rangle\varphi \mid E\varphi.$

The intended reading of $\langle\leq\rangle\varphi$ is "φ is true in a world that is at least as good as the current world", while $\langle<\rangle\varphi$ says that "φ is true in a world that is strictly better than the current world. ". In addition, the auxiliary *existential modality $E\varphi$* says that "there is a world where φ is true". As usual, we write $[\leq]\varphi$ for the defined universal modality $\neg\langle\leq\rangle\neg\varphi$, and we use $[<]$ and U for the duals of $\langle<\rangle\varphi$ and E, respectively. Combinations of these modalities can capture a wide variety of binary preference statements comparing propositions, witness the cited literature.

The interpretation of this modal language over our models is as follows:

Definition 16. *Truth conditions for the atomic propositions and Boolean combinations are standard. Modalities are interpreted like this:*

- $\mathfrak{M}, s \models \langle\leq\rangle\varphi$ *iff for some t wih $s \preceq t$, $\mathfrak{M}, t \models \varphi$.*
- $\mathfrak{M}, s \models \langle<\rangle\varphi$ *iff for some t with $s \prec t$, $\mathfrak{M}, t \models \varphi$.*
- $\mathfrak{M}, s \models E\varphi$ *iff for some world t in W, $\mathfrak{M}, t \models \varphi$.*

The defined modalities use the obvious universal versions of these clauses. For concreteness, we state the standard calculus to come out of this.

Theorem 17. *Modal betterness logic is completely axiomatized by*

1. *the system **S4** for the preference modality,*
2. *the system **S5** for the universal modality,*
3. *the connecting law $U\varphi \to [\preceq]\varphi$,*
4. *three axioms for the strict betterness modality: cf. [16].*

4.2 Special features of preference

Next we briefly survey three special logical features of preference structure that go beyond standard modal logic of pre-orders, and that will eventually turn out to be of interest to deontics as well.

Lifting to generic preferences. While betterness relates specific objects or worlds, preference is often used generically for comparing different *kinds* of things. Ever since [142], logicians have also studied preferences $P(\varphi, \psi)$ between propositions, viewed as properties of worlds, or of objects.

There is not one such notion, but many, that can be defined by a *lift* of the betterness order among worlds to sets of worlds, cf. [65], [16], [94]. For instance, compare your next moves in a game, identified with the set of outcomes that they lead to. Which move is 'better' depends on the criterion chosen: maybe we want the one with the highest possible outcome, or the one with the highest minimally guaranteed outcome, etcetera.

Such options are reflected in various quantifier combinations for the lifting. In particular, von Wright had a $\forall\forall$-type preference between sets P, Q:

$$\forall x \in P \ \forall y \in Q\colon\ x \preceq y.$$

A simpler, but also useful example is the modal $\forall\exists$-type

$$\forall x \in P \ \exists y \in Q\colon\ x \preceq y.$$

This says that for any P-world, there is a Q-world which is at least as good as that ψ-world. In the earlier game setting, this stipulation would say that the most preferred moves have the highest maximal outcomes. This ubiquitous $\forall\exists$ generic preference can be defined in the above modal preference language, using the universal modality ranging over all worlds:

$$P^{\forall\exists}(\varphi, \psi) := U(\psi \to \langle\leq\rangle\varphi).$$

This generic preference $P\varphi\psi$ satisfies the usual properties for preference, reflexivity and transitivity: for instance, $P\varphi\psi$ and $P\psi\chi$ imply $P\varphi\chi$.[20]

Ceteris paribus clauses. Unlike plausibility, preference ordering seldom comes in pure form: the comparison between alternatives is often entangled with other considerations. Again, games provide an example. Usually, players do not compare moves via the sets of all their possible outcomes, but rather, they compare the *most plausible* outcomes of their moves. This is the so-called *normality sense* of ceteris paribus preference: we do not compare all the φ and ψ-worlds, but only the 'normal ones' in some relevant sense. This belief restriction, observed by many authors, will return in our discussion of doxastic entanglement of preference in Section 8.

But there are also other natural senses of taking a ceteris paribus clause. It was noticed already in [142] that there is also an 'equality sense' of preference, involving

[20] Other stipulationss lead to other generic preferences. This proliferation may be a problem (e.g., 'doing what is best' depends on one's stipulation as to 'best'), but there is no consensus in the literature. A logical approach at least helps make the options clear.

a hidden assumption of *independence*. In that case, one only make comparisons between worlds where some things or issues are held constant, in terms of giving the same truth values to some specified set of atomic propositions, or complex formulas. The logic of equality-based preference is axiomatized and analyzed in detail in [16].

Richer preference languages. Modal languages are just one step on a ladder of formalisms for analyzing reasoning practices. It has been claimed that richer languages are needed to faithfully render basic preference notions, cf. [38] on first-order preferences among objects, [60] on first-order languages of social choice, [20] on hybrid modal preference languages for defining backward induction solutions in games, the hybrid modal language of 'desire' and 'freedom' for decision making in [64], or the modal fixed-point languages for games used in [13]. Though we will mainly use modal formalisms to make the essential points to follow, we will mention the relevance of such richer preference formalisms occasionally.

5 World based dynamics of preference change

Now let us look at how given preferences can change. Intuitively, there are many acts and events that can have such an effect. Perhaps the purest form is a radical *command* by some moral authority to do something. This makes the worlds where we act better than those where we do not, cf. [146]: at least, if we 'take' the order as a legitimate instruction, and change our evaluation accordingly, overriding any preferences that we ourselves might have had. Technically, this dynamics will change a current betterness relation in a model. This can be studied entirely along the lines already developed here for information dynamics.[21]

5.1 Betterness change

[26] is a first systematic study of betterness change using methods from dynamic-epistemic logic. The running example in their approach is a weak 'suggestion' $\sharp\varphi$ that a proposition φ be the case. This relatively modest ordering change leaves the set of worlds the same, but it removes any preferences that the agent might have had for $\neg\varphi$-worlds over φ-worlds among these.[22]

The main general point to note here is that events with evaluative import can act as triggers that change some current betterness relation on worlds. In particular, a suggestion $\sharp\varphi$ leads to the following model change:

[21] Of earlier treatments, we mention [135], based on [139].

[22] Similar operations have come up recently in logical treatments of relevant alternatives theories in epistemology, when modeling changes in what is considered relevant to making or evaluating a knowledge claim. Cf. [77], [24].

Definition 18. *Given any modal preference model* $(\mathfrak{M}, s)$*, the* suggestion upgrade $(\mathfrak{M}\sharp\varphi, s)$ *has the same domain, valuation, and actual world as* $(\mathfrak{M}, s)$*, but the new preference relations are now*

$$\preceq_i^* = \preceq_i - \{(s,t) \mid \mathfrak{M}, s \models \varphi \textit{ and } \mathfrak{M}, t \models \neg\varphi\}$$

In preference models $\mathfrak{M}$, a matching dynamic modality is interpreted as:

$$(\mathfrak{M}, s) \models [\sharp\varphi]\psi \quad \text{iff} \quad \mathfrak{M}_{\sharp\varphi}, s \models \psi$$

Again, complete dynamic logics exist (cf. [26]). The reader will find it useful to scrutinize the key recursion law for preferences after suggestion.[23]

Theorem 19. *The dynamic preference logic of suggestion is completely axiomatized by the following principles:*

1. $\langle\sharp\varphi\rangle p \ \leftrightarrow \ p$
2. $\langle\sharp\varphi\rangle\neg\psi \ \leftrightarrow \ \neg\langle\sharp\varphi\rangle\psi$
3. $\langle\sharp\varphi\rangle(\psi \wedge \chi) \ \leftrightarrow \ (\langle\sharp\varphi\rangle\psi \wedge \langle\sharp\varphi\rangle\chi)$
4. $\langle\sharp\varphi\rangle\langle\leq\rangle\psi \ \leftrightarrow \ (\neg\varphi \wedge \langle\leq\rangle\langle\sharp\varphi\rangle\psi) \vee (\varphi \wedge \langle\leq\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi))$
5. $\langle\sharp\varphi\rangle E\psi \ \leftrightarrow \ E\langle\sharp\varphi\rangle\psi$

Similar completeness results are presented in [94] for dynamic logics that govern many other kinds of normative action, such as the 'strong commands' corresponding to our earlier radical plausibility upgrade. Following this instruction, deontically, the agent incorporates the wish of some over-riding authority.

5.2 Deriving changes in defined preferences

This is an analysis of betterness change and modal statements about it local to specific worlds. But it also applies to the earlier lifted *generic preferences.* As an illustration, consider the ∀∃-lift defined earlier:

Fact 20. *The following equivalence holds for generic* ∀∃ *preference:*

$$\langle\sharp A\rangle P^{\forall\exists}(\varphi,\psi) \quad \textit{iff} \quad P^{\forall\exists}(\langle\sharp A\rangle\varphi, \langle\sharp A\rangle\psi) \wedge P^{\forall\exists}((\langle\sharp A\rangle\varphi \wedge A), (\langle\sharp A\rangle\psi \wedge A)).$$

We omit the simple calculation for this outcome. Similar results may be obtained for other set liftings such as Von Wright's ∀∀-version.

Finally, the recursive style of dynamic analysis presented here also applies to various forms of ceteris paribus preference.

[23] Technically, the simplicity of this law reflects the clear analogy between our universal preference modality and the earlier doxastic notion of safe belief.

5.3 General formats for betterness change

Behind our specific examples of betterness change, there lies a much more general theory that works for a wide class of triggering events that change betterness or evaluation order. One widely applicable way of achieving greater generality uses programs from *propositional dynamic logic PDL*.

For instance, suggesting that φ is defined by the program:

$$\sharp\varphi(R) := (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; R; ?\varphi).$$

where R is the given input relation, while the operations $?\varphi$ test whether the relevant proposition φ, or related ones, hold. In particular, the disjunct $(?\varphi; R; ?\varphi)$ means that we keep all old betterness links that run from φ-worlds to φ-worlds.

This definition is equivalent in PDL to the more compact program expression

$$\sharp\varphi(R) := (?\neg\varphi; R) \cup (R; ?\varphi).$$

Again we keep all old R-links, except for those that ran from φ-worlds to $\neg\varphi$-worlds.

Likewise, our plausibility changers for belief revision can be defined in this format. For instance, the earlier 'radical upgrade' is defined by

$$\Uparrow\varphi(R) := (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; \top; ?\varphi)$$

Here the constant symbol $\top$ denotes the universal relation that holds between any two worlds. This reflects the original meaning of this transformation: all φ-worlds become better than all $\neg\varphi$-worlds, whether or not they were better before, and within these two zones, the old ordering remains.[24]

Given any PDL program definition of the above sort, one can automatically write recursion laws for the complete dynamic logic of its induced model change, cf. [26] for the precise algorithm. As an illustration, here is the straightforward computation for suggestions:

$$\begin{aligned}
\langle\sharp\varphi\rangle\langle R\rangle\psi \quad &\leftrightarrow \quad \langle(?\neg\varphi; R) \cup (R; ?\varphi)\rangle\langle\sharp\varphi\rangle\psi \\
&\leftrightarrow \quad \langle ?\neg\varphi; R\rangle\langle\sharp\varphi\rangle\psi \vee \langle R; ?\varphi\rangle\langle\sharp\varphi\rangle\psi \\
&\leftrightarrow \quad (\neg\varphi \wedge \langle R\rangle\langle\sharp\varphi\rangle\psi) \vee \langle R\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi).
\end{aligned}$$

[24] Conservative upgrades can be dealt with in a similar way. As commands, these leave the agent more of her original preferences: so, differences with radical commands will show up in judgments of 'conditional betterness', as discussed in the literature on conditional obligation: see [67].

For alternative general formats of ordering change supporting our sort of dynamic logics, we refer to the 'priority update' with event models in [10], the order merge perspective of [21], as well as the still more general 'dynamic dynamic logic' of [55].

In our view, the practical and theoretical theoretical variety of ordering changes for plausibility and preference is not a nuisance, but a feature. It matches the wealth of evaluative actions that we encounter in daily life.

6 Reason-based preferences

Primitive betterness relations among worlds or objects reflect what are called 'intrinsic preferences'. But very often, our preferences have an underlying structure, and we compare according to criteria: our preferences are then reason-based, or 'extrinsic'. In this section we develop the latter view, that has motivations in linguistic Optimality Theory, cf. [112], and belief revision based on entrenchment, cf. [116]. This view also occurs in reason-based deontic logic, cf. [48], [56] and [81], as we shall see in Section 9.

A simplest illustration of our approach, that suffices for many natural scenarios, starts with linear orders of relevant properties that serve as criteria for determining our evaluation of objects or worlds.

6.1 Priority based preference

The following proposal has many ancestors, among which we mention the treatment in [49], [116]. We follow [38], that starts from a given primitive ordering among propositions ('priorities' among properties of objects or worlds), and then derives a preference among objects themselves.

Definition 21. *A* priority sequence *is a finite linear sequence of formulas written as follows:* $C_1 \gg C_2 \cdots \gg C_n \quad (n \in \mathbb{N})$*, where the* C_m *come from a language describing objects, with one free variable* x *in each* C_m*.*

Definition 22. *Given a priority sequence and objects* x *and* y*, Pref*(x, y) *is defined lexicographically: at the first property* C_i *in the given sequence where* x, y *have a different truth value,* $C_i(x)$ *holds, but* $C_i(y)$ *fails.*

The logic of this framework is analyzed in [38], while applications to deontic logic are developed in [25]. Still, this is only one of many ways of deriving a preference from a priority sequence. A good overview of existing approaches is found in [36].

6.2 Pre-orders

In general, comparison order need not be connected, and then the preceding needs a significant generalization. This was done, in a setting of social choice and belief merge, in the seminal paper [2], which we adapt slightly here to the notion of 'priority graphs', based on the treatment in [54], [95].

The following definitions contain a free parameter for a *language* L that can be interpreted in the earlier modal betterness models $\mathfrak{M}$. For simplicity only, we will take this to be a simple propositional language of properties.

Definition 23. *A* priority graph $\mathscr{G} = \langle P, < \rangle$ *is a strictly partially ordered set of propositions in the relevant language of properties* L.

Here is how one derives a betterness order from a priority graph:

Definition 24. *Let* $\mathscr{G} = \langle P, < \rangle$ *be a priority graph, and* $\mathfrak{M}$ *a model in which the language* L *defines properties of objects. The* induced betterness relation $\preceq_{\mathscr{G}}$ *between objects or worlds is defined as follows:*

$$y \preceq_{\mathscr{G}} x := \forall P{\in}\mathscr{G}((Py \to Px) \vee \exists P'{<}P(P'x \wedge \neg P'y)).$$

Here, in principle, $y \preceq_{\mathscr{G}} x$ requires that x has every property in the graph that y has. But there is a possibility of 'compensation': if y has P while x does not, this is admissible, provided there is some property P' with higher priority in the graph where x does better: x has P' while y lacks it. Clearly, this stipulation subsumes the earlier priority sequences: linear priority graphs lead to lexicographic order.

One can think of priority graphs of propositions in many ways that are relevant to this paper. In the informational realm, they are hierarchically ordered information sources, structuring the evidence for agents' beliefs. In the normative realm, they can stand for complex hierarchies of laws, or of norm givers with relative authority.

6.3 Static logic and representation theorem

In what follows, we immediately state a crucial technical property of this framework, cf. [49], [95].

Theorem 25. *Let* $\mathfrak{M} = (W, \preceq, V)$ *be any modal preference model, without constraints on its relation. The following two statements are equivalent:*

(a) *The relation* $y \preceq x$ *is a reflexive and transitive order,*
(b) *There is a priority graph* $\mathscr{G} = (P, <)$ *such that,*
for all worlds $x, y \in W$, $y \preceq x$ *iff* $y \preceq_{\mathscr{G}} x$.

This representation theorem says that the general logic of derived extrinsic betterness orderings is still just that of pre-orders. But it also tells us that any intrinsic pre-order can be rationalized as an extrinsic reason-based one by adding structure without disturbing the base model as it is.

6.4 Priority dynamics and graph algebra

Now, we have a new locus for more fine-grained preference change: the family of underlying reasons, which brings its own logical structure. For linear priority sequences, relevant changes involve the obvious operations $[^+C]$ of adding a new proposition C to the right, $[C^+]$ of adding C to the left, and various functions $[-]$ dropping first, last or intermediate elements of a priority sequence. [38] give complete dynamic logics for these. Here is one typical valid principe:

$$[^+C]Pref(x,y) \leftrightarrow Pref(x,y) \vee (Eq(x,y) \wedge C(x) \wedge \neg C(y))$$

This set of natural operations for changing preferences becomes even richer in the realm of priority graphs, due to their possibly non-linear structure. However, in this setting an elegant mathematical alternative arises, in terms of merely two fundamental operations that combine arbitrary graphs:

- $\mathscr{G}_1;\mathscr{G}_2$ adding a graph to another in top position
- $\mathscr{G}_1\|\mathscr{G}_2$ adding two graphs in parallel.

One can think of this as the obvious counterparts of ‘sequential’ versus ‘parallel’ composition. Here the very special case where one of the graphs consists of just one proposition models simple update actions.

This graph calculus has been axiomatized completely in [2] by algebraic means, while [54] presents a further modal-style axiomatization. We display its major modal principles here, since they express the essential recursion underlying priority graph dynamics. Here is one case where, as mentioned earlier, a slight language extension is helpful: in what follows, the proposition letter n is a ‘nominal’ from hybrid logic denoting one single world.

$$\langle \mathscr{G}_1\|\mathscr{G}_2\rangle^{\leq} n \;\leftrightarrow\; \langle \mathscr{G}_1\rangle^{\leq} n \wedge \langle \mathscr{G}_2\rangle^{\leq} n.$$

$$\langle \mathscr{G}_1\|\mathscr{G}_2\rangle^{<} n \;\leftrightarrow\; (\langle \mathscr{G}_1\rangle^{<} n \wedge \langle \mathscr{G}_2\rangle^{\leq} n) \vee (\langle \mathscr{G}_1\rangle^{\leq} n \wedge \langle \mathscr{G}_2\rangle^{<} n).$$

$$\langle \mathscr{G}_1;\mathscr{G}_2\rangle^{\leq} n \;\leftrightarrow\; (\langle \mathscr{G}_1\rangle^{\leq} n \wedge \langle \mathscr{G}_2\rangle^{\leq} n) \vee \langle \mathscr{G}_1\rangle^{<} n.$$

$$\langle \mathscr{G}_1;\mathscr{G}_2\rangle^{<} n \;\leftrightarrow\; (\langle \mathscr{G}_1\rangle^{\leq} n \wedge \langle \mathscr{G}_2\rangle^{<} n) \vee \langle \mathscr{G}_1\rangle^{<} n.$$

These axioms reduce complex priority relations to simple ones, after which the whole language reduces to the modal logic of weak and strict atomic betterness orders. In particular, this modal graph logic is decidable.

Thus, we have shown how putting reasons underneath agents' preferences (or, for that matter, their beliefs) admits of precise logical treatment, while still supporting the systematic dynamics that we are after.

7 A two-level view of preference

Now we have two ways of looking at preference: one through intrinsic betterness order on modal models, the other through priority structure inducing extrinsic betterness orders. One might see this as calling for a reduction from one level to another, but instead, *combining* the two perspectives seems the more attractive option, as providing a richer modeling tool for preference-driven agency.

7.1 Harmony of world order and reasons

In many cases, the two modeling levels are in close harmony, allowing for easy switches from one to the other (cf. [91]):

Definition 26. *Let α: $(\mathscr{G}, A) \to \mathscr{G}'$, with $\mathscr{G}$, $\mathscr{G}'$ priority graphs, and let A be a new proposition. Let σ be a map from $(\preceq, A)$ to $\preceq'$, where $\preceq$ and $\preceq'$ are betterness relations over worlds. We say that α* induces *σ, if always:*

$$\sigma(\preceq_{\mathscr{G}}, A) \quad = \quad \preceq_{\alpha(\mathscr{G},A)}$$

Here are two results that elaborate the resulting harmony between two levels for our earlier major betterness transformers:

Fact 27. *Taking a suggestion A is the map induced by the priority graph update $\mathscr{G}\|A$. More precisely, the following diagram commutes:*

$$\begin{tikzcd}
\langle \mathscr{G}, < \rangle \arrow[r, "\|A"] \arrow[d] & \langle (\mathscr{G}\|A), < \rangle \arrow[d] \\
\langle W, \preceq \rangle \arrow[r, "\sharp A"] & \langle W, \sharp A(\preceq) \rangle
\end{tikzcd}$$

For a second telling illustration of such harmony in terms of our earlier themes, consider a priority graph $(\mathscr{G}, <)$ with a new proposition A added on top. The logical dynamics at the two levels is now correlated as follows:

Fact 28. *Placing a new proposition A on top of a priority graph $(\mathscr{G}, <)$ induces the radical upgrade operation $\Uparrow A$ on possible worlds ordering models. More precisely, the following diagram commutes:*

$$\begin{tikzcd}
\langle \mathscr{G}, < \rangle \arrow[r, "A;\mathscr{G}"] \arrow[d] & \langle (A;\mathscr{G}), < \rangle \arrow[d] \\
\langle W, \preceq \rangle \arrow[r, "\Uparrow A"] & \langle W, \Uparrow A(\preceq) \rangle
\end{tikzcd}$$

Thus the two kinds of preference dynamics dovetail well: [94] has details.

7.2 Correlated dynamics

There are several advantages to working at both levels without reductions. For a start, not all natural operations on graphs have matching betterness transformers at all. An example from [95] is *deletion* of the topmost elements from a given priority graph. This syntactic operation of removing criteria is not invariant for replacing graph arguments by other graphs inducing the same betterness order, and hence it is a genuine extension of preference change.

But also conversely, there is no general match. Not all PDL-definable betterness changers from Section 5.3 are graph-definable. In particular, not all PDL transformers preserve the basic order properties of reflexivity and transitivity guaranteed by priority graphs. For a concrete illustration, consider the program

$?A; R$: 'keep the old relation only from where A is true'.

This change does not preserve reflexivity of an order relation R, because the $\neg A$-worlds now have no outgoing relation arrows any more.[25]

All this argues for a more general policy of modeling both intrinsic and extrinsic preference for agents, with reasons for the latter encoded in priority graphs that are an explicit part of the modeling.

Still, one might think that intrinsic betterness relations merely reflect an agent's raw feelings or prejudices. But the intrinsic-extrinsic contrast is relative, not absolute. If I obey the command of a higher moral authority, I may acquire an extrinsic preference, whose reason is obeying a superior. But for that higher agent, the same preference may be intrinsic: "The king's whim is my law". This observation suggests a further theme: transitioning from one perspective to the other.

[25] Intuitively, the operation $?A; R$ amounts to a refusal to make betterness comparisons at worlds that lack property A. Though idiosyncratic, this seems a bona fide mind change for an agent.

7.3 Additional dynamics: language change

Technically, intrinsic betterness can become extrinsic through a dynamics that has been largely outside the scope of dynamic-epistemic logic so far, that of *language change*. One mechanism here is the proof of the earlier representation result stated in Theorem 25. It partitions the given betterness pre-order into clusters, and if these are viewed as new relevant reasons or criteria, the resulting strict order of clusters is a priority graph inducing the given order. This may look like mere formal rationalization, but in practice, one often observes agents' preferences between objects, and then postulates reasons for them. A relevant source is the notion of 'revealed preference' from the economics literature: cf. [79].

Thus, our richer view of preference also suggests a new kind of dynamics beyond what we have considered so far. In general, reasons for given preferences may have to come from some other, richer language than the one that we started with: we are witnessing a dynamic act of *language creation*.[26]

8 Combining evaluation and information

We have now completed our exposition of information dynamics as well as preference dynamics, which brought its own further topics. What must have become abundantly clear is that there are strong formal similarities in the logic of order and order change in the two realms. We have not even enumerated all of these similarities, but, for instance, all of our earlier ideas and results about reason-based preference also make sense when analyzing evidence-based belief.

This compatibility helps with the next natural step we must take. As we said right at the start of this paper, the major agency systems of information and evaluation do not live in isolation: they interact all the time. A rational agent can process information well in the sense of proof or observation, but is also 'reasonable' in a broader sense of being guided by goals. This *entanglement* of knowledge, belief, and preference shows in many specific settings. We will look at a few cases, and in particular, their impact on the dynamics of preference change.[27] Though we will mainly discuss how information dynamics influences preference and deontic notions, the opposite influence is equally real. In particular, information flow depends on *trust* and *authority*: which are clearly deontic notions.[28]

[26] For a study of language change in the setting for belief revision, cf. [108].

[27] For a more general discussion, we refer to [109].

[28] Following Wittgenstein, Brandom (cf. [34]) has even argued that language use can only be fully understood in terms of commitments that carry rights and obligations.

8.1 Generic preference with knowledge

In Section 4.2, we defined one basic generic preference as follows:

$Pref^{\forall\exists}(\psi, \varphi) := U(\psi \to \langle\leq\rangle\varphi)$.

This refers to possibilities in the whole model, including even those that an agent might know to be excluded. [26] defend this scenario in terms of 'regret', but still, there is also a reasonable intuition that preference only runs among situations that are epistemically possible.

This suggests the entangled notion that, for any ψ-world that is *epistemically accessible* to agent a in the model, there is a world which is at least as good where φ is true. This can be written with an epistemic modality:

$$Pref^{\forall\exists}(\varphi, \psi) ::= K_a(\psi \to \langle\leq\rangle\varphi). \qquad (K_{bett})$$

But this is not yet what we are after, since we want the 'better world' to be epistemically accessible itself. [92] shows how this cannot be defined in a simple combined language of knowledge and betterness, and that instead, a richer preference formalism is needed with a new *intersection modality* for epistemic accessibility and betterness. The latter entangled notion can be axiomatized, and it also supports a dynamic logic of preference change as before.[29]

8.2 Generic preference with belief

Issues of entanglement become even more appealing with generic preference and belief, where the two relational styles of modeling were very similar to begin with. Again, we might start with a mere combination formula

$$Pref^{\forall\exists}(\varphi, \psi) ::= B_a(\psi \to \langle\leq\rangle\varphi). \qquad (B_{bett})$$

This says that, among the most plausible worlds for the agent, for any ψ-world, there exists a world which is at least as good where φ is true.[30]

Again, this seems not quite right in many cases, since we often want the better worlds relevant to preference to stay inside the most plausible part of the model, being 'informational realists' in our desires. To express this, we again need a stronger

[29] An alternative approach would be to impose *additional modal axioms* that require betterness alternatives to be epistemic alternatives via frame correspondence. However, this puts constraints on our dynamic operations transforming models that we have not investigated. We leave this alternative line as a topic for further investigation.

[30] One might also think here of using a *conditional belief* $B^{\psi}\langle\leq\rangle\varphi$, but to us, this seems an intuitively less plausible form of entanglement.

merge of the two relations by intersection. The key clause for a corresponding new modality then reads like a 'wishful safe belief':

$$\mathfrak{M}, s \models H\varphi \text{ iff } \textit{for all } t \textit{ with both } s \leq t \textit{ and } s \preceq t, \mathfrak{M}, t \models \varphi.$$

As before, the static and dynamic logic of this entangled notion yield to the general dynamic-epistemic methodology explained in earlier sections.

8.3 Other entanglements of preference and normality

Entangled versions of plausibility and betterness abound in the literature. For instance, [33] has models $\mathfrak{M} = (W, \leq_P, \leq_N, V)$ with W a set of possible worlds, V a valuation function and $\leq_P$, $\leq_N$ two transitive connected relations $x \leq_P y$ ('y is as good as x) and $x \leq_N y$ ('y is as normal as x). These models support an operator of *conditional ideal goal* (IG):

$$\mathfrak{M} {\models} IG^{\psi}\varphi \;\text{ iff }\; Max(\leq_P, Max(\leq_N, Mod(\psi))) \subseteq Mod(\varphi)$$

This says that the best of the most normal ψ worlds satisfy φ. Such entangled notions are still expressible in the modal systems of this chapter.

Fact 29. $IG^{\psi}\varphi ::= U(\psi \wedge \neg\langle B^{<}\rangle\psi) \wedge \neg\langle<\rangle(\psi \wedge \neg\langle B^{<}\rangle\psi) \rightarrow \varphi)$.[31]

Following up on this tradition in agency studies in computer science, the paper [87] defines the following entangled notion of preference:

Definition 30. $\mathfrak{M} \models Pref^{*}(\varphi, \psi)$ *iff for all* $w' \in Max(\leq_N, Mod(\psi))$*, there exists* $w \in Max(\leq_N, Mod(\varphi))$ *such that* $w' <_P w$.

This reflects the earlier-mentioned 'ceteris paribus' sense of preference, where one compares only the normal worlds of the relevant kinds.[32] Intriguingly, a source of similar ideas is the semantics of expressions like "want" and "desire" in natural language, cf. [128], [74], [37].

The preceding notions are similar to our earlier one with an intersection modality, but not quite. They only compare the two most plausible parts for each proposition.

We give no deeper analysis of all these entangled notions here, but as one small appetizer, we note that we are still within the bounds of this paper.

Fact 31. *$Pref^{*}$ is definable in a modal doxastic preference language.*

[31] Here, $B^{<}$ is an earlier-mentioned modality of *strong belief* that we do not define.

[32] This makes sense in epistemic game theory, where 'rationality' means comparing moves by their most plausible consequences according to the player's beliefs and then choosing the best.

8.4 Preference change and belief revision

As we have observed already, our treatment of the statics and dynamics of belief and preference shows many similarities. It is an interesting test, then, if the earlier dynamic logic methods transfer to entangled notions of preference. Intuitively, entangled preferences can change because of two kinds of trigger: evaluative acts like suggestions or commands, and informative acts changing our beliefs. As a positive illustration, we quote one result from [91]:

Theorem 32. *The dynamic logic of the above intersective preference H is axiomatizable, with the following essential recursion axioms:*

1. $\langle \sharp A\rangle\langle H\rangle\varphi \leftrightarrow (A \wedge \langle H\rangle(A \wedge \langle \sharp A\rangle\varphi)) \vee (\neg A \wedge \langle H\rangle\langle \sharp A\rangle\varphi)$.
2. $\langle \Uparrow A\rangle\langle H\rangle\varphi \leftrightarrow (A \wedge \langle H\rangle(A \wedge \langle \Uparrow A\rangle\varphi)) \vee (\neg A \wedge \langle H\rangle(\neg A \wedge \langle \Uparrow A\rangle\varphi)) \vee$
$(\neg A \wedge \langle bett\rangle(A \wedge \langle \Uparrow A\rangle\varphi))$.
3. $\langle A!\rangle\langle H\rangle\varphi \leftrightarrow A \wedge \langle H\rangle\langle A!\rangle\varphi$.

Having intersection modalities may not be all that is needed, though, since there may also be *entangled triggering events* that do not easily reduce to purely informational or purely evaluative actions.[33]

Trade-offs between preference change and information change. Finally, as often in logic, distinctions can get blurred through redefinition. For instance, sometimes, the same scenario may be modeled either in terms of preference change, or as information change. Two concrete examples of such redescription are “Buying a House” in [38] and “Visit by the Queen” in [88]. Important though it is, we leave the study of precise connections between different representations of dynamic entangled scenarios to another occasion.

9 Deontic reasoning, changing norms and obligations

Our analysis of information and preference can itself be viewed as a study of normative discourse and reasoning. However, in this section, we turn to explicit deontic scenarios, and take a look at some major issues concerning obligations and norms from the standpoint of dynamic systems for preference change.[34]

[33] For an analogy, see the question scenarios involving conversational triggers for parallel information and issue change in [19].

[34] Our treatment largely follows the papers [25], [17].

Perhaps the most immediate concrete task at hand is charting the large variety of deontic actions in daily life that affect normative betterness orderings. These normative triggers range from commands to promises and permissions. We will not undertake such a survey here, but the examples in this paper will hopefully convince the reader that a dynamic action perspective on deontic issues is natural, and that much can be done with the tools presented here. Instead, we consider four general topics that have roots in the deontic literature.

9.1 Unary and dyadic obligation on ordering models

Our static logics heavily relied on binary ordering relations. In fact, deontic logic was first with this approach, building on observations from ethics that the deontic notions of obligation, permission and prohibition can be naturally made sense of in terms of an *ideality ordering* $\preceq$ on possible worlds. Here is an early quote from [101], found in [48], p.6.

> " [...] to assert that a certain line of conduct is [...] absolutely right or obligatory, is obviously to assert that more good or less evil will exist in the world, if it is adopted, than if anything else be done instead."

In this line, the pioneering study [67] interpreted dyadic obligations of the type 'it is obligatory that φ under condition ψ' on semantic models like ours, using a notion of maximality as in our study of belief:

$$\mathfrak{M}, s \models \mathbf{O}(\varphi \mid \psi) \Longleftrightarrow Max(||\psi||_{\mathfrak{M}}) \subseteq ||\varphi||_{\mathfrak{M}}$$

Depending on the properties of the relation $\preceq$, different deontic logics are obtained here: [67] starts with a $\preceq$ which is only reflexive, moving then to total pre-orders. This is of course the same idea that has also emerged in conditional logic, belief revision, and the linguistic semantics of generic expressions.[35] Variations of this modeling have given rise to various preference-based semantics of deontic logic: see [134] for an overview.

In this light, our paper has taken up an old idea in the semantics of deontic reasoning, and then added some recent themes concerning preference: criterion-based priority structure, dynamics of evaluative acts and events, and extended logical languages making these explicit. This seems a natural continuation of deontic logic, while also linking it up with developments in other fields.

[35] One deontic criticism of this account has been that it made conditional obligation lack the property of antecedent strengthening: [132]. This, however, makes perfect sense in our view, as it reflects precisely the non-monotonicity inherent in the dynamics of information change, where the most ideal worlds can change during update.

9.2 Reasons and dynamics in deontic paradoxes

The dynamic emphasis in this paper on changes and their triggering events has thrown fresh light on the study of information and preference-based agency. Deontic logic proves to be no exception, if we also bring in our treatment of reason-based preference – as we shall see with a few examples.

The Gentle Murder scenario from [46], p.194, is a classic of deontic logic that illustrates the basic problem of 'contrary-to-duty' obligations (CTDs).

Example 33. *"Let us suppose a legal system which forbids all kinds of murder, but which considers murdering violently to be a worse crime than murdering gently. [...] The system then captures its views about murder by means of a number of rules, including these two:*

1. *It is obligatory under the law that Smith not murder Jones.*
2. *It is obligatory that, if Smith murders Jones, Smith* [does so] *gently."*

The priority format of Section 6.1, even just linear sequences, can represent this scenario in a natural way. Recall that a linear priority sequence $P_1, \ldots, P_n$ combines bipartitions $\{\mathcal{I}(p_i), -\mathcal{I}(p_i)\}$ of the domain of discourse S. Moving towards the right direction of the sequence, ever more atoms p_i are falsified. In a deontic reading, this means that, the more we move towards the right side of the sequence, the more violations hold of morally desirable properties.

Concretely, in the Gentle Murder scenario, the result is two classes of ideality: one class I_1 in which Smith does not murder Jones, i.e., $\mathsf{I}_1 := \neg m$; and another I_2 in which either Smith does not murder Jones or he murders him gently, i.e., $\mathsf{I}_2 := \neg m \vee (m \wedge g)$. The relevant priority sequence $\mathcal{B}$ has $\mathsf{I}_2 \prec \mathsf{I}_1$. Such a sequence orders the worlds via its induced relation $\preceq_{\mathcal{B}}^{IM}$ in three clusters. The most ideal states are those satisfying I_1, worse but not worst states satisfy $\mathsf{V}_1 := \neg \mathsf{I}_1$ but at the same time I_2, and, finally, the worst states satisfy $\mathsf{V}_2 := \neg \mathsf{I}_2$.

With this representation, we can take the scenario one step further.

Example 34. *Consider the priority sequence for Gentle Murder from the preceding Example: $\mathcal{B} = (\mathsf{I}_1, \mathsf{I}_2)$. We can naturally* restrict *$\mathcal{B}$ to an occurrence of the first violation by intersecting all formulas in the sequence with V_1. Then the first proposition becomes a contradiction, distinguishing no worlds. The best among the still available worlds are those with $Max^{+}(\mathcal{B}^{\mathsf{V}_1}) = \mathsf{I}_2 \wedge \mathsf{V}_1$. A next interesting restriction is $\mathcal{B}^{\mathsf{V}_2}$, which represents what the original priority sequence prescribes under the assumption that also the CTD obligation "kill gently" has been violated. In this case we end up in a set of states that are all equally bad.*

This brief sketch may suffice to show our approach provides a simple perspective on the deontic robustness of norms and laws viewed as *CTD* structures: they can still function when transgressions have taken place.[36]

Other major puzzles in the deontic literature, such as the Chisholm Paradox, are given similar reason-based representations in [17].

9.3 Typology of change at two levels

We have shown how two-level structure of preference provides a natural medium for modeling deontic notions. Likewise, it yields a rich account of deontic changes. In Section 7, we developed a theory of both informational and evaluative changes, either directly on possible world order, or on priority structure underlying such orders. This also makes sense here.

As an illustration, we add a temporal twist to the above deontic scenario, by 'dynamifying' Gentle Murder.

Example 35. *We start with a priority sequence $\mathcal{B} = (\neg m)$. This current deontic state of affairs generates a total pre-order where all $\neg m$-states are above all m-states: "It is obligatory under the law that Smith not murder Jones". Now, we refine this order so as to introduce the sub-ideal obligation to kill gently: "it is obligatory that, if Smith murders Jones, Smith murders Jones gently". In other words, we want to model the process of refining legal codes, by introducing a contrary-to-duty obligation.*

Intuitively, this change can happen in one of two ways:

1. *We refine the given betterness ordering 'on the go' by requesting a further bipartition of the violation states, putting the $m \wedge g$-states above the $m \wedge \neg g$-states. This can be seen as the successful execution of a command of the sort "if you murder, then murder gently".*

2. *We introduce a new law 'from scratch', where $m \to g$ is now explicitly formulated as a class of possibly sub-ideal states. This can be seen as the enactment of a new priority sequence $(\neg m, m \to g)$.*[37]

The example illustrates how a *CTD* sequence can be dynamically created either by uttering a sequence of commands stating what ought to be the case in a sub-ideal situation, or by enacting a new priority sequence.

[36] Representing *CTD* structures as finite chains of properties already occurs informally in [48]. The first formal account is in [57], where an elegant Gentzen calculus is developed for handling formulae of the type $\varphi_1 @ \ldots @ \varphi_n$ with @ a connective representing a sort of 'sub-ideality' relation. It is an interesting open problem if such a proof calculus can be embedded in the modal logics of this chapter.

[37] We have encountered this before, since $m \to g$ is equivalent to $\neg m \vee (m \wedge g)$.

But in this setting, Theorem 27 from Section 7 applies: in terms of betterness among worlds, the two instructions amount to the same thing! In other words, in this scenario, the same deontic change can be obtained both by refining the order dictated by a given law, and by enacting a new law.

Of course, this is just a start, and not everything is smooth application. Our discussion of two-level dynamics in Section 7.2 also suggests that some well-known changes in laws, such as *abrogation* (a counterpart to the earlier operation of 'graph deletion') have no obvious counterpart at the pure worlds level.

9.4 Norm change

The preceding discussion leads up to a more general theme of global dynamics. The problem of *norm change* has recently gained attention from researchers in deontic logic, legal theory, as well as multi-agent systems.

Approaches to norm change fall into two groups. In syntactic approaches—inspired by legal practice—norm change is an operation performed directly on the explicit provisions in the code of the normative system [58], [59], [31]. In semantic approaches, however, norm change follows deontic preference order (cf. also [6]). Our initial betterness dynamics on models belonged to the latter group, but our priority methods tie it to the former.[38]

More drastic changes of norms and moral codes can be modeled, too, in our framework, using the calculus of priority graphs that we have sketched in Section 6. For details, we refer again to [17].

9.5 Entangled changes

Finally, as observed already in Section 8 on entanglement (cf. [87] for a deontic discussion), the dynamic logic connection allows for a unified treatment of *t*wo kinds of change that mix harmoniously in deontic reasoning: information change given a fixed normative order, and evaluation change modifying such an order. Deontic scenarios can have deeply intertwined combinations of obligation, knowledge and belief (cf. [94]). Some sophisticated moral scenarios in [106] include natural dynamic issues that we have ignored here, such as the subtle, but real difference between 'knowing one's duty' versus 'having a duty to know'.

Many further dynamic deontic themes can be analyzed along the above lines. We refer to [25], [17] for a detailed treatment of the Chisholm Paradox, and concrete ways in which priority graph calculus models norm change.

[38] The bridge here is our earlier analysis: obligations defined via ideality and maximality are special kinds of classifications of an Andersonian-Kangerian type.

10 Further directions

Collecting points from earlier sections, here are a few further directions where deontic logic meets with current trends in dynamic logics of agency.

10.1 Language, speech acts, and agency

Events that drive information or preference change are often *speech acts* of telling, asking, and so on. Natural language has a sophisticated repertoire of speech acts with a deontic flavour (commanding, promising, allowing, and so on) that invite further logical study, taking earlier studies in meta-ethics and Speech Act Theory (cf. [119]) to the next level. In particular, such studies will also need a more fine-grained account of the *multi-agency* in dynamic triggers, that has been ignored in this chapter. For instance, things are said by someone to someone, and their uptake depends on relations of authority or trust. Likewise, promises, commands, or permissions are given by someone to someone, and their normative effect depends in subtle ways on who does, and is, what. [148] is a pioneering study of this fine-structure of normative action using dynamic-epistemic logic.

10.2 Multi-agency and groups

A conspicuous turn in studies of information dynamics has been a strong emphasis on social scenarios with more than one agent: [12], [121], [7], [66]. After all, the natural paradigm for language use is communication between different agents, a major historical source for logic is argumentation between different parties, social behaviour is kept in place by mutual expectations, and so on.

In the logics for knowledge, belief, and preference of this paper, part of this multi-agent turn can be represented by mere iteration of single-agent modalities, as in a's knowing that b does, or does not, knows some fact. But the next stage is the introduction of *groups* as agents, where logics have been devised for notions such as 'common knowledge' or 'distributed knowledge' in groups, and likewise for beliefs (cf. [45], [100]), or the group-level preferences underlying Social Choice Theory (cf. [44]). All these logics also have dynamic-epistemic extensions in the style of this chapter, although systematic extensions to, say, social choice or judgement aggregation remain to be developed.

The social turn is highly relevant to deontic logic. From the start, deontic notions and morality seems all about *others*: my duties are usually toward other people, my norms come from outside sources: my boss, or a lawgiver.[39] In principle, the methods

[39] This social aspect has been clearly acknowledged by computer scientists working on multi-agent

of this paper can deal with social multi-agent structure in deontic settings, though much remains to be investigated. For instance, it is easy to interpret informational iterations $K_a K_b p$, but what, for instance, is the meaning of an iterated obligation $O_a O_b p$? And beyond this, what is a group-based 'common obligation': is this more like common belief, or like a demand for joint action of the group? Other relevant issues are the entanglement of informational and evaluative acts for groups: cf. [73], [84], [83], and [75] on morality as held together by social expectations such as trust. An account of deontically relevant actions for groups will also have to include new operations reminiscent of social choice, such as *belief merge* and *preference merge*, where the priority structures of Section 6 may find a new use: this time, as a model for institutions (cf. [61]).

10.3 Games and dependent behaviour

Multi-agency is tied together not just by social knowledge or beliefs, but also by dependent individual and collective *action*. Thus, logics of agency have close connections with game theory ([125], [12]) and the study of strategic behavior and its equilibria. In deontic practice, dependent action is crucial (think of sanctions or rewards), and games are a congenial paradigm. Many topics in this paper suggest game-theoretic analogies. We already saw how belief-entangled set lifting is crucial to player's choices and their rationality, making preference logics a natural tool in the analysis of games (cf. [118], [39], [13]). Conversely, ideas from game theory have entered deontic logic, witness the use of game solution methods as moral deliberation procedures in [96]. One might even argue that dependent behaviour is the source of morality, and in that sense, games would be the really natural next stage after the single-episode driven dynamic logics of this paper.

10.4 Temporal perspective

Games are one longer-term activity, but deontic agency involves many different processes, some even infinite. The general logical setting here are temporal logics (cf. [45], [110]) where new phenomena come to the fore. Deontics and morality is not just about single episodes, but about action and interaction over time. Early work in deontic logic already used temporal logics: cf. the pioneering dissertation [43]) where events happen in infinite histories, and obligations come and go. Likewise, in the multi-agent community, logics have been proposed for preferences between complete histories, and planning behaviour leading toward most desired histories (cf. [98], [122]). Such temporal logics mesh well with dynamic-epistemic logics (cf.

systems: cf. [99], [145], and [114].

[14]), with an interesting role for *protocols* as a new object of study, i.e., available procedures that both provide and constrain available actions for reaching goals. Plans and protocols have a clear normative dimension as well, and one would wish to incorporate them into the preference dynamics of this paper.

10.5 Syntax and fine-structure

Most dynamic logics for agency, whether about information dynamics or evaluation dynamics, are semantic in nature. The states changed by the process are semantic models. Still, in philosophical logic, there is a continuing debate about the right representation of *information.* Semantic information, though common to many areas, including decision theory and game theory, is coarse-grained, identifying logically equivalent propositions, suppressing the very act of logical analysis as an information-producing process. Zooming in on the latter, agents engage in many activities, such as inference, memory retrieval, introspection, or other forms of 'awareness management' that require a more fine-grained notion of information, closer to syntax. Several dynamic logics of this kind have been proposed in recent years (cf. [18], [82], [137]).

The same issues of grain level for information make sense in the deontic realm. For instance, our priority graphs were syntactic objects than get manipulated by insertions, deletions, permutations, and the like. But also, deontic logic has its own counterpart to the epistemological problem of 'omniscience'. My moral obligations to you cannot reasonably be based on my foreseeing every consequence of my commitments. I owe you careful deliberation, not omniscience.[40] Here too, there is a need for more fine-grained dynamic representations, closer to deontic syntax.

10.6 Numerical strength

While the main theme of this chapter is qualitative approaches, there are also numerical approaches to preferences, employing utilities (cf. [115], [133]) or more abstract 'grades' for worlds (cf. [127]). Dynamic ideas work in this setting, too, witness the modal logic with graded modalities indicating the strength of preference in [5], which also defines product update for numerical plausibility models. A stream-lined version in [90] uses propositional constants q_a^m saying that agent a assigns the current world a value of at most m. Our earlier ordering models, both for plausibility and for preference, now get numerical graded versions, with more finely-grained statements of strength of belief and of preference. Dynamic updates can still be defined,

[40] Likewise, citizens are supposed to know the law, but they need not be professional lawyers in seeing every relevant deductive consequence.

where we assign values to actions or events, using numerical stipulations in terms of 'product update' from the cited references.[41] More complex numerical evaluation uses *utility* as a fine-structure of preference, and its dynamics can also be dealt with in this style: cf. [90], [93].

While the technical details of these approaches are not relevant here, systems like this do address two issues that seem of great deontic relevance. One is the possibility of comparing not just worlds qua preference, but also actions, making sense of the principled distinction in ethics between outcome-oriented and deontological views of obligations and commitments. The other major feature is that we can now study the more quantitative logic of *how much good* an action does, and the extent to which we can *improve* current situations by our actions.

10.7 Probability

Another obvious quantitative addition to our analysis would be *probability.* Probabilities measure strengths of beliefs, thereby providing fine-structure to the plausibility orderings that we have worked with. But they can also indicate information that we have about a current process, or a reliability we assign to our observation of a current event.[42] Finally, the numerical factors in probability theory also allow us to mix and weigh various factors in the entangled versions of preference and deontic notions that we have discussed in Section 8. A striking entangled notion is *expected value* in probability theory, whose definition mixes beliefs and evaluation. A treatment of such notions in our current framework remains a desideratum.

11 Appendix: relevant strands in the literature

The themes of this paper have a long history, with many proposals in the literature for combining and 'dynamifying' preferences, beliefs, and obligations. In addition to those cited already, here are some other relevant lines of work.

Computation and agency. [99] is a pioneering study of deontics from a dynamic viewpoint, reducing deontic logics to suitable dynamic logics. In the same tradition, [98] takes the deontic logic/dynamic logic interface a step further, studying 'free choice permission' with a new dynamic logic where preferences can hold between actions. Completeness theorems for this enriched semantics then result for several systems. [113] provide a dynamified logic of permission that builds action policies for

[41] The resulting dynamic logic of numerical evaluation can be axiomatized in the same recursive style as the qualitative systems that we have discussed in this paper.

[42] See [15] for a rich dynamic-epistemic logic of reasoning with and updating probability.

agents by adding or deleting transitions. [40] reduces an extension of van der Meyden's logic to PDL, yielding an EXPTIME decision procedure, and showing how PDL can deal with agents' policies. Preference semantics has also been widely used in AI tasks: e.g., [144] gives a preference-based semantics for goals in decision theory. This provides criteria for verifying the design of goal-based planning strategies, and a new framework for knowledge-level analysis of planning systems. [78] studies commonsense normative reasoning, arguing that techniques of non-monotonic logic provide a better framework than the usual modal treatments. The paper has applications to conflicting obligations and conditional obligations. [87] propose a logic of desires whose semantics contains two ordering relations of preference and normality, and then interpret "in context A, I desire B" as 'the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds', providing a new entanglement of preference and normality.

Semantics of natural language. In a line going back to [127], [139] presents an update semantics for default rules, locating their meaning in the way in which they modify expectation patterns. This is part of a general program of 'update semantics' for conditionals and other key expressions in natural language. [135] use ideas from update semantics to formalize deontic reasoning about obligations. In their view, the meaning of a normative sentence resides in the changes it brings about in the 'ideality relations' of agents to whom a norm applies. [149] uses a simple dynamic update logic to formalize natural language imperatives of the form $FIAT\ \varphi$, which can be used in describing the search for solutions of planning problems. [97] extends the update semantic analysis of imperatives to include third person and past tense imperatives, while also applying it to the notion of free choice permission. [107] outlines a preference-based account of communication, which brings the dynamics of changing obligations for language users to the fore. [147] distinguishes the illocutionary acts of commanding from the perlocutionary acts that affect preferences of addressees, proposes a new dynamic logic which combines preference upgrade and deontic update, and discusses some deontic dilemmas in this setting.

Philosophical logic. The philosophical study of agency has many themes that are relevant to this paper, often inspired by topics in epistemology or by the philosophy of action. In a direction that is complementary to ours, with belief change as a starting point, [70] identifies four types of changes in preference, namely revision, contraction, addition and subtraction, and shows that they satisfy plausible postulates for rational changes. The collection [63] brings together the latest approaches on preference change from philosophy, economics and psychology. Following Hansson's work, [1] defines minimal preference change in the spirit of AGM framework and characterises minimal contraction by a set of postulates. A linear time algorithm is proposed for computing preference changes. In addition, going far beyond what we have discussed in this paper, Hansson has written a series of seminal papers combining ideas from preference logic and deontic logic, see e.g. [69], [68] and [72].

Rational choice theory. Preference is at the heart of decision and rational choice. In recent work at the interface of preference logic, philosophy, and social science, themes from our chapter such as reason-based and belief-entangled preference have come to the fore, with further lines of their own. [42] and [41] point out that, though existing decision theory gives a good account of how agents make choices given their preferences, issues of where these essential preferences come from and how they can change are rarely studied.[43] The authors propose a model in which agents' preferences are based on 'motivationally salient properties' of alternatives, consistent sets of which can be compared using a 'weighing relation'. Two intuitive axioms are identified in this setting that precisely characterize the property-based preference relations. Starting from similar motivations, [102] studies reason-based preference in more complex doxastic settings, drawing on ideas from similarity-based semantics for conditional logic. Essentially, preference results here from agents' comparing two worlds, one having some property and the other lacking it, close to their actual world, and comparing these based on relevant aspects of utility. The framework supports extensive analysis in modal logic, including illuminating results on frame correspondence and axiomatization. [103] gives an extension of this approach to preference in the presence of quantifiers, while [104] makes a link between these preference models and deontic logic. A detailed comparison of the two mentioned recent approaches with the one in this paper remains to be undertaken.

12 Conclusion

We have shown how dynamic logics of agency can deal with information, criteria, and preference change. In doing so, we obtained a suggestive framework for the analysis of deontic notions that connects many strands in the literature on agency.

References

[1] N. Alechina, F. Liu, and B. Logan. Minimal preference change. In D. Grossi, O. Roy, and H. Huang, editors, *Proceedings of the 4th International Workshop on Logic, Rationality and Interaction (LORI 2013)*, volume 8196 of *FoLLI-LNCS*, pages 15–26. Springer, 2013.

[43] These are of course precisely the two main topics of this paper: cf. also [94].

[2] H. Andréka, M. Ryan, and P-Y. Schobbens. Operators and laws for combining preferential relations. *Journal of Logic and Computation*, 12:12–53, 2002.

[3] L. Åqvist. *Introduction to Deontic Logic and the Theory of Normative Systems.* Naples: Bibliopolis, 1987.

[4] L. Åqvist. Deontic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 605–714. Dordrecht: Kluwer, 1994.

[5] G. Aucher. A combined system for update logic and belief revision. Master's thesis, MoL-2003-03. ILLC, University of Amsterdam, 2003.

[6] G. Aucher, D. Grossi, A. Herzig, and E. Lorini. Dynamic context logic. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 15–26. Springer, 2009.

[7] A Baltag, Z. Christoff, J.U. Hansen, and S. Smets. Logical models of informational cascades. In J. van Benthem and F. Liu, editors, *Logic Across the University: Foundations and Application*, pages 405–432. College Publications, London, 2013.

[8] A. Baltag, N. Gierasimczuk, and S. Smets. Belief revision as a truth tracking process. In *Proceedings of Theoretical Aspects of Rationality and Knowledge(TARK 2011)*, 2011.

[9] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pages 43–56, 1998.

[10] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In M. Wooldridge G. Bonanno, W. van der Hoek, editor, *Logic and the Foundations of Game and Decision Theory*, volume 3 of *Texts in Logic and Games.* Amsterdam: Amsterdam University Press, 2008.

[11] A. Baltag, S. Smets, and J. Zvesper. Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.

[12] J. van Benthem. *Logical Dynamics Information And Interaction.* Cambridge University Press, 2011.

[13] J. van Benthem. *Logic in Games.* The MIT Press, 2014.

[14] J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *Journal of Philosophical Logic*, 38(5):491–526, 2009.

[15] J. van Benthem, J. Gerbrandy, and B. Kooi. Dynamic update with probabilities. *Studia Logica*, 93(1):67–96, 2009.

[16] J. van Benthem, P. Girard, and O. Roy. Everything else being equal: A modal logic approach for ceteris paribus preferences. *Journal of Philosophical Logic*, 38(1):83–125, 2009.

[17] J. van Benthem, D. Grossi, and F. Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[18] J. van Benthem and M. Martínez. The stories of logic and information. In J. van Ben-

them and P. Adriaans, editors, *Handbook of Philosophy of Information.* Amsterdam: Elsevier, 2008.

[19] J. van Benthem and S. Minica. Toward a dynamic logic of questions. In X. He, J. F. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 27–41. Springer, 2009.

[20] J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals and solution concepts in games. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 61–77. Uppsala Philosophical Studies 53, 2006.

[21] J.van Benthem. Belief update as social choice. In P. Girard, O. Roy, and M. Marion, editors, *Dynamic Formal Epistemology*, pages 151–160. Springer, Dordrecht, 2006.

[22] J.van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic*, 17:129–156, 2007.

[23] J.van Benthem. *Modal Logic for Open Minds.* Stanford: CSLI Publications, 2010.

[24] J.van Benthem. Talking about knowledge. To appear in C. Baskent, L. Moss and R. Ramanujam, eds., Volume in Honor of Rohit Parikh, Outstanding Contributions to Logic, Springer, Dordrecht, 2014.

[25] J.van Benthem, D. Grossi, and F. Liu. Deontics = betterness + priority. In G. Governatori and G. Sartor, editors, *Deontic Logic in Computer Science, 10th International Conference, DEON 2010*, volume 6181 of *LNAI*, pages 50–65. Springer, 2010.

[26] J.van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic*, 17:157–182, 2007.

[27] J.van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

[28] J.van Benthem and S. Smets. Dynamic logic of belief change. To appear in H. van Ditmarsch, J. Halpern, W. van der Hoek & B. Kooi, eds., Handbook of Logics of Knowledge and Belief, College Publications, London, 2014.

[29] P. Blackburn, J. van Benthem, and F. Wolter. *Handbook of Modal Logic.* Elsevier, 2007.

[30] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Cambridge: Cambridge University Press, 2001.

[31] G. Boella, G. Pigozzi, and L. van der Torre. Normative framework for normative system change. In P.Decker, J.Sichman, C.Sierra, and C.Castelfranchi, editors, *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 169–176, 2009.

[32] C. Boutilier. *Conditional Logics for Default Reasoning and Belief Revision.* PhD thesis, University of Toronto, 1992.

[33] C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68:87–154, 1994.

[34] R. Brandom. *Making it Explicit: Reasoning, Representing, and Discursive Commit-*

ment. Harvard University Press, 1994.

[35] J. Burgess. Basic tense logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 89–133. Dordrecht: D. Reidel, 1984.

[36] S. Coste-Marquis, J. Lang, P. Liberatore, and P. Marquis. Expressive power and succinctness of propositional languages for preference representation. In *Proceedings of the 9th International Conference on Principles of Knowledge Representation and Reasoning (KR 2004).* Menlo Park, CA: AAAI Press, 2004.

[37] S. Dandelet. Partial desires, blinkered beliefs. Manuscript, January 22, 2014.

[38] D. de Jongh and F. Liu. Preference, priorities and belief. In T.Grune-Yanoff and S.O. Hansson, editors, *Preference Change: Approaches from Philosophy, Economics and Psychology*, Theory and Decision Library, pages 85–108. Springer, 2009.

[39] C. Dégremont. *The Temporal Mind. Observations on the Logic of Belief Change in Interactive Systems.* PhD thesis, ILLC, University of Amsterdam, 2010.

[40] S. Demri. A reduction from DLP to PDL. *Journal of Logic and Computation*, 15:767–785, 2005.

[41] F. Dietrich and C. List. A reason-based theory of rational choice. *Nous*, 47(1):104–134, 2013.

[42] F. Dietrich and C. List. Where do preferences come from? *International Journal of Game Theory*, 42(3):613–637, 2013.

[43] J. van Eck. *A System of Temporally Relative Modal and Deontic Predicate Logic and its Philosophical Applications.* PhD thesis, University of Groningen, 1981.

[44] U. Endriss. Applications of logic in social choice theory. In *Proceedings of the 12th International Workshop on Computational Logic in Multiagent Systems (CLIMA-2011)*, volume 6814 of *LNAI*, pages 88–91. Springer-Verlag, July 2011. Extended abstract corresponding to an invited talk.

[45] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge.* Cambridge, MA: The MIT Press, 1995.

[46] J. Forrester. Gentle murder, or the adverbial samaritan. *Journal of Philosophy*, 81:193–197, 1984.

[47] B. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438, 1972.

[48] B. van Fraassen. Values and the heart's command. *The Journal of Philosophy*, 70(1):5–19, 1973.

[49] N. Friedman and J. Halpern. Plausibility measures: A user's guide. In *Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence (UAI 95)*, pages 175–184, 1995.

[50] N. Friedman and Joseph Y. Halpern. Modeling belief in dynamic systems, part I: Foundations. *Artificial Intelligence*, 95(2):257–316, 1997.

[51] N. Friedman and Joseph Y. Halpern. Modeling belief in dynamic systems. part II: Revision and update. *Journal of Artificial Intelligence Research*, 10:117–167, 1999.

[52] D. Gabbay, J.Horty, R. van der Meyden, and L. van der Torre. Handbook on Deontic Logic and Normative Systems, 2014. To appear.

[53] J. Gerbrandy. *Bisimulation on Planet Kripke.* PhD thesis, ILLC, University of Amsterdam, 1999.

[54] P. Girard. *Modal Logics for Belief and Preference Change.* PhD thesis, Stanford University, 2008.

[55] P. Girard, J. Seligman, and F. Liu. General dynamic dynamic logic. In Thomas Bolander, Torben Braüner, Silvio Ghilardi, and Lawrence S. Moss, editors, *Advances in Modal Logic*, pages 239–260. College Publications, 2012.

[56] L. Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5(2):113–134, 2000.

[57] G. Governatori and A. Rotolo. Logic of violations: A gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 3:193–215, 2005.

[58] G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment. part 1: Revision and defeasible theories. In R. van der Meyden and L. van der Torre, editors, *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON 2008)*, volume 5076 of *LNAI*, pages 3–18. Springer, 2008.

[59] G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment. part 2: Temporalised defeasible logic. In G. Boella, G. Pigozzi, M. P. Singh, and H. Verhagen, editors, *Proceedings of the 3rd International Workshop on Normative Multiagent System (NorMAS 2008), Luxembourg, Luxembourg, July 14-15, 2008*, pages 112–127, 2008.

[60] U. Grandi and U. Endriss. First-order logic formalisation of arrow's theorem. In *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI-2009)*, volume 5834 of *LNAI*, pages 133–146. Springer-Verlag, October 2009. Also presented at DGL-2009.

[61] D. Grossi. *Desigining Invisible Hancuffs. Formal Investigations in Institutions and Organizations for Multi-Agent Systems.* PhD thesis, Utrecht University, 2007. SIKS Dissertation Series 2007-16.

[62] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[63] T. Grune-Yanoff and S.O. Hansson, editors. *Preference Change: Approaches from Philosophy, Economics and Psychology.* Theory and Decision Library. Springer, 2009.

[64] M. Guo and J. Sliegman. Making choices in social situations. In *Logic and Interactive Rationalityok*, pages 176–202. ILLC, University of Amsterdam, 2011.

[65] J.Y. Halpern. Defining relative likelihood in partially-ordered preferential structure. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.

[66] P. G. Hansen and V. F. Hendricks. *Infostorms: How to Take Information Punches and Save Democracy.* Copernicus Books / Springer, 2014.

[67] B. Hansson. An analysis of some deontic logics. *Nous*, 3:373–398, 1969.

[68] S.O Hansson. Defining 'good' and 'bad' in terms of 'better'. *Notre Dame of Journal of Formal Logic*, 31:136–149, 1990.

[69] S.O Hansson. Preference-based deontic logic. *Journal of Philosophical Logic*, 19:75–93,

1990.

[70] S.O Hansson. Changes in preference. *Theory and Decision*, 38:1–28, 1995.

[71] S.O Hansson. Preference logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 319–393. Dordrecht: Kluwer, 2001.

[72] S.O Hansson. *The Structure of Values and Norms*. Cambridge: Cambridge University Press, 2001.

[73] G. den Hartog. *Wederkerige Verwachtingen [Mutual Expectations]*. PhD thesis, University of Amsterdam, 1985.

[74] I. Heim. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9(3):183–221, 1992.

[75] W. H. Holliday. Dynamic testimonial logic. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality, and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 161–179. Springer, 2009.

[76] W. H. Holliday. *Knowing what follows: epistemic closure and epistemic logic*. PhD thesis, Stanford University, 2012.

[77] W. H. Holliday. Epistemic closure and epistemic logic i: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, 2014.

[78] J. Horty. Deontic logic as founded on nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9:69–91, 1993.

[79] D. Houser and R. Kurzban. Revealed preference, belief, and game theory. *Economics and Philosophy*, 16:99–115, 2002.

[80] T. Icard III, E. Pacuit, and Y. Shoham. Joint revision of beliefs and intentions. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 572 – 574. AAAI Publications, 2010.

[81] F. Jackson. On the semantics and logic of obligation. *Mind*, XCIV:177–195, 1985.

[82] M. Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, 2006.

[83] J. Konkka. Funk games: Approaching collective rationality. E-Thesis, University of Helsinki, 2000.

[84] B. Kooi and A. Tamminga. Conflicting obligations in multi-agent deontic logic. In John-Jules Ch. Meyer and Lou Goble, editors, *Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science*, volume 4048 of *LNCS*, pages 175–186. Springer, 2006.

[85] P. Lamarre. S4 as the conditional logic of nonmonotonicity. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 357–367, 1991.

[86] P. Lamarre and Y. Shoham. Knowledge, certainty, belief, and conditionalisation abbreviated version. In *KR'94*, pages 415–424, 1994.

[87] J. Lang, L. van der Torre, and E. Weydert. Hidden uncertainty in the logical representation of desires. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pages 189–231, 2003.

[88] J. Lang and L. van der Torre. From belief change to preference change. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-2008)*, pages 351–355, 2008.

[89] D. Lewis. *Counterfactuals.* Oxford: Blackwell, 1973.

[90] F. Liu. Dynamic variations: Update and revision for diverse agents. Master's thesis, MoL-2004-05. ILLC, University of Amsterdam, 2004.

[91] F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity.* PhD thesis, ILLC, University of Amsterdam, 2008.

[92] F. Liu. Logics for interaction between preference and belief. Manuscript, Department of Philosophy, Tsinghua University, 2009.

[93] F Liu. Preference change: A quantitative approach. *Studies in Logic*, 2(3):12–27, 2009.

[94] F. Liu. *Reasoning about Preference Dynamics*, volume 354 of *Synthese Library.* Springer, 2011.

[95] F. Liu. A two-level perspective on preference. *Journal of Philosophical Logic*, 40:421–439, 2011.

[96] L. O. Loohuis. Obligations in a responsible world. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 251–262. Springer, 2009.

[97] R. Mastop. *What Can You Do? Imperative Mood in Semantic Theory.* PhD thesis, ILLC, University of Amsterdam, 2005.

[98] R. van der Meyden. The dynamic logic of permission. *Journal of Logic and Computation*, 6:465–479, 1996.

[99] J-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[100] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence.* Cambridge: Cambridge University Press, 1995.

[101] G. E. Moore. *Principia Ethica.* Cambridge University Press, 1903.

[102] D. Osherson and S. Weinstein. Preference based on reasons. *The Review of Symbolic Logic*, 5:122–147, 3 2012.

[103] D. Osherson and S. Weinstein. quantified preference logic. Manuscript, October 5, 2012.

[104] D. Osherson and S. Weinstein. Deontic modality based on preference. Manuscript, July 4, 2014.

[105] S. van Otterloo. *A Strategic Analysis of Multi-agent Protocols.* PhD thesis, Liverpool University, UK, 2005.

[106] E. Pacuit, R. Parikh, and E. Cogan. The logic of knowledge based on obligation. *Synthese*, 149:311–341, 2006.

[107] X. Parent. Remedial interchange, contrary-to-duty obligation and commutation. *Journal of Applied Non-Classical Logics*, 13(3/4):345–375, 2003.

[108] R. Parikh. Beliefs, belief revision, and splitting languages. In J. Ginzburg, L. Moss, and M. de Rijke, editors, *Logic, Language and Computation*, volume 2, pages 266–278. Center for the Study of Language and Information Stanford, CA, 1999.

[109] R. Parikh, L. O. Loohuis, and C. Baskent. Epistemic norms. To appear in Dov Gabbay, John Horty, Ron van der Meyden and Leon van der Torre eds, Handbook on Deontic Logic and Normative Systems, College Publications, London, 2011.

[110] R. Parikh and R. Ramanujam. A knowledge-based semantics of messages. *Journal of Logic, Language and Information*, 12:453–467, 2003.

[111] J.A. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216, 1989.

[112] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar.* Oxford: Blackwell, 2004.

[113] R. Pucella and V. Weissmann. Reasoning about dynamic policies. In *Proceedings FoSSaCS-7*, Lecture Notes in Computer Science 2987, pages 453–467, 2004.

[114] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. San Mateo, CA: Morgan Kaufmann, 1991.

[115] N. Rescher. Notes on preference, utility, and cost. *Synthese*, 16:332–343, 1966.

[116] H. Rott. Basic entrenchment. *Studia Logica*, 73:257–280, 2003.

[117] H. Rott. Shifting priorities: Simple representations for 27 iterated theory change operators. In H. Langerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 359–384. Uppsala Philosophical Studies 53, 2006.

[118] O. Roy. *Thinking before Acting: Intentions, Logic and Rational Choice.* PhD thesis, ILLC, University of Amsterdam, 2008.

[119] J. R. Searle and D. van der Veken. *Foundations of Illocutionary Logic.* Cambridge: Cambridge University Press, 1985.

[120] K. Segerberg. The basic dynamic doxastic logic of AGM. In M.-A. Williams and H. Rott, editors, *Frontiers in Belief Revision*, pages 57–84. Kluwer Academic Publishers, 2001.

[121] J. Seligman, F. Liu, and P. Girard. Facebook and the epistemic logic of friendship. In Burkhard. C. Schipper, editor, *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 229–238, 2013.

[122] M. Sergot. $(C+)^{++}$: An action language for modelling norms and institutions. Technical Report 8, Department of Computing, Imperial College, London, 2004.

[123] C. Shi. Logics of evidence-based belief and knowledge. Master's thesis, Tsinghua University, 2014.

[124] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence.* Cambridge, MA: The MIT Press, 1988.

[125] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations.* Cambridge: Cambridge University Press, 2008.

[126] J. Snyder. Product update for agents with bounded memory. Manuscript, Department of Philosophy, Stanford University, 2004.

[127] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics II*, pages 105–134. Dordrecht: Kluwer, 1988.

[128] R. Stalnaker. *Inquiry.* Cambridge University Press: Cambridge, United Kingdom, 1984.

[129] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(2):133–163, 1996.

[130] R. Stalnaker. Extensive and strategic forms: Games and models for games. *Research in Economics*, 53:293–319, 1999.

[131] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199, 2006.

[132] Y.-H. Tan and L. van der Torre. How to combine ordering and minimizing in a deontic logic based on preferences. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems, DEON '96: The 3rd International Workshop on Deontic Logic in Computer Science*, pages 216–232, 1996.

[133] R.W. Trapp. Utility theory and preference logic. *Erkenntnis*, 22:301–339, 1985.

[134] L. van der Torre. *Reasoning about Obligations: Defeasibility in Preference-based Deontic Logic.* PhD thesis, Rotterdam, 1997.

[135] L. van der Torre and Y.-H. Tan. An update semantics for deontic reasoning. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems*, pages 73–90. Amsterdam: IOS Press, 1999.

[136] H. Van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic.* Berlin: Springer, 2007.

[137] F. R. Velazquez-Quesada. Inference and update. *Synthese*, 169:283–300, 2009.

[138] F. Veltman. *Logics for Conditionals.* PhD thesis, University of Amsterdam, 1985.

[139] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[140] G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[141] G. H. von Wright. A note on deontic logic and derived ogligation. *Mind*, 65:507–509, 1956.

[142] G. H. von Wright. *The Logic of Preference.* Edinburgh: Edinburgh University Press, 1963.

[143] G. H. von Wright. A new system of deontic logic. In *Danish Yearbook of Philosophy*, pages 173–182. Museum Tusculanum Press, Copenhagen, 1964.

[144] M. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of the National Conference on Artificial Intelligence*, pages 698–703, 1991.

[145] M. Wooldridge. *Reasoning about Rational Agents.* Cambridge, MA: The MIT Press, 2000.

[146] T. Yamada. Acts of commands and changing obligations. In K. Inoue, K. Satoh, and F. Toni, editors, *Proceedings of the 7th Workshop on Computational Logic in Multi-Agent Systems (CLIMA VII)*, 2006. Revised version appeared in LNAI 4371, pages 1-19, Springer-Verlag, 2007.

[147] T. Yamada. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, 165(2):295–315, 2008.

[148] T. Yamada. Scorekeeping and dynamic logics of speech acts. Manuscript, Hokkaido University, 2010.

[149] B. Zarnic. Imperative change and obligation to do. In K. Segerberg and R. Sliwinski, editors, *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Aqvist*, pages 79–95. Uppsala philosophical studies 51. Uppsala: Department of Philosophy, Uppsala University, 2003.

Received July 2014

The Formalization of Practical Reasoning: Problems and Prospects

Richmond H. Thomason
Philosophy Department, University of Michigan

Abstract

Deontic logic, as traditionally conceived, provides only a deductive theory that constrains the states or possible worlds within which an agent should try to remain. As such, it only encompasses a small part of practical reasoning, which in general is concerned with selecting, committing to, and executing plans. In this article I try to frame the general challenge that is presented to logical theory by the problem of formalizing practical reasoning, and to survey the existing resources that might contribute to the development of such a formalization. I conclude that, while a robust, adequate logic of practical reasoning is not yet in place, the materials for developing such a logic are now available.

Keywords: Formal Practical Reasoning, Imperative Inference, Agent Architecture, Desires , Intentions.

1 The challenge of formalizing practical reasoning

Practical reasoning is deliberation. It is reasoning about what to do. We do it all the time. Any day in our life will provide us with hundreds of examples of thinking about what to do. But it has been remarkably difficult to produce a comprehensive, adequate theory of practical reasoning. Part of the difficulty is that the topic is studied by different disciplines, each of these has something important to contribute, and it is unusual to find a study of practical reasoning that brings all of these perspectives together.

I will begin by considering examples of practical reasoning. (I suspect that the range of examples is broader than many people might imagine.) I will then propose a rationale for classifying these examples, and canvass the disciplines that have something useful to say about the reasoning.

A more or less comprehensive inventory of examples will provide an idea of what an adequate account of practical reasoning might look like. In the remainder of the paper, I try to say something about the challenges that an approach that begins to do justice to the subject would have to address.

1.1 Some Examples

All too many published discussions of practical reasoning — even book-length discussions — cover only a very small part of the territory. For that reason, it's vital to begin with a broad range of examples.

Example 1. Ordering a meal at a restaurant.

The deliberating agent sits down in a restaurant and is offered a menu. Here, the problem is deciding what to eat and drink. Suppose that the only relevant factors are price and preferences about food. Even for a moderately sized menu and wine list, the number of possible combinations is over $400,000$. It would be very unlikely for an ordinary human being to work out a total preference ordering for each option. In fact, even though the decision will probably involve weighing preferences about food and drink against preferences about cost, the reasoning might well produce a decision without appealing to a general rule for reconciling these preferences.

Example 2. Deciding what move to make in a chess game.

In chess, an individual action needs to be evaluated in the context of its continuations. There is no uncertainty about the current state or the immediate consequences of actions, but much uncertainty about moves that the opponent might make. The *search space* (i.e., the number of possible continuations) is enormous — on the order of 10^{43}. Determining the value of positions involves conflicting criteria (e.g. positional advantages versus numerical strength); these conflicts must be resolved in comparing the value of different positions. In tournament chess, deliberation time is limited. These somewhat artificial constraints combine to concentrate the reasoning on exploration of a search space. Perhaps because of this, the reasoning involved in chess has been intensively investigated by psychologists and computer scientists, and influenced the classical work on search algorithms in AI; see [48].

Example 3. Savage's omelet.

In [44][pp. 13–15], Leonard Savage describes the problem as follows.

> Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must either be used for the omelet or wasted altogether,

> lies beside the bowl. You must decide what to do with the unbroken egg. . . . you must decide between three acts only, namely, to break it into the bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection.

This problem involves preferences about the desired outcomes, as well as risk, in the form of a positive probability that the egg is spoiled. The problem is to infer preferences over actions. The outcomes are manifest and involve only a few variables, the preferences over them are evident, and the probabilities associating each action with an outcome can be easily estimated. In this case, the reasoning reduces to the calculation of an expected utility.

Example 4. Designing a house.

This example is less obviously practical; it is possible for an architect to design a house without thinking much about the actions that will go into building it, leaving this to the contractor.[1] However, an architect's design becomes the builder's goals, and I would maintain that inferring goals is a form of practical reasoning. The reasoning combines constraint satisfaction and optimization, where again conflicts between competing desiderata may need to be resolved. Any real-life architect will also use *case-based reasoning*, looking in a library of known designs for one that is relevant, and modifying a chosen example to suit the present purpose.

Example 5. Deciding how to get to the airport.

This is a planning problem; the agent a has an inventory of actions, knows their preconditions and effects, knows the relevant features of the current state, and has as its goal a state in which a is at the airport. In its simplest form, the problem is to find a sequence of actions that will transform the current state into a state that satisfies the goal. Planning, or means-end reasoning, is one of the most intensively studied forms of reasoning in AI. The earliest planning algorithms made many simplifying assumptions about the planning situation and the conditions that a satisfactory plan must meet; over the years, sophisticated planning algorithms have been developed that depend on fewer of these assumptions and so can be used in a variety of realistic settings.[2]

[1] Of course, a good design has to take into account how to build a house, in order to make sure that the design is feasible.

[2] See, for instance, [39]. For the airport problem in particular, see [34].

Example 6. Cracking an egg into a bowl.

This is a case in which most of of us do the action automatically, with hardly any conscious reasoning. Probably most people can't remember the circumstances under which they learned how to do it. But the activity is complex: there are many ways to get it wrong. This example was proposed as a benchmark problem in the formalization of common-sense reasoning. The literature on this problem shows that the reasoning is surprisingly complicated, and it presupposes much common-sense knowledge; see, for instance, [45]. This example is different from the previous ones in that the solution to the reasoning problem is acted out; the reasoning must engage motor systems, and it depends on these systems for grasping and manipulating objects according to plan. For obvious reasons, Savage ignored this part of the omelet problem.

Example 7. Playing table tennis.

Unlike chess, table tennis is a game in which practical reasoning has to be *online*; engaged in complex, real-time activities involving the perceptual and motor systems. For a novice, the reasoning may be exhausted by the need to keep the ball in play; experts may be able to engage in tactical reasoning. But there is no time to spare for reflection; the reasoning needs to be thoroughly connected to the ongoing process of play.

Example 8. Playing soccer.

Soccer is like table tennis, but with the added dimension of teamwork and the need to recognize and execute plays. This task was selected as a benchmark problem in robotics, and has been extensively studied. See, for instance, [54, 40, 3].

Example 9. Typing a message.

Typing an email message, composing it as you go along, starts perhaps with a general idea of what to say. The reasoning that produced a rough idea of the content may have taken place reflectively, but once composition has begun, several reasoning processes are engaged simultaneously, and have to be coordinated. The general idea of what to say has to be packaged in linguistic form, and this form has to be rendered by motor actions at the keyboard. For a skilled typist composing a straightforward message, these complex, practical tasks are combined and executed very quickly, perhaps at the rate of 70 words per minute. For this to happen, the interface between high-level linguistic reasoning and motor skills has to be very robust.

Example 10. Factory scheduling.

The factory scheduler has to produce, say on a daily basis, a sequence of manufacturing operations for each order to be processed that day, and a schedule allocating times and machines to these operations. This problem is notorious for the difficulty of the reasoning; it involves horrible combinatorics, uncertainty, limited time for reflection, and the resolution of many conflicting desiderata. Among the goals cited by [17] are (1) meeting order dates, (2) minimizing work-in-process time, (3) maximizing allocation of factory resources, and (4) minimizing disruption of shop activity.

Part of the interest of this example lies in the difference in scale between this problem and Savage's omelet problem. It is not clear that there is any way to construct a single, coherent utility function for the task, by reconciling the four desiderata mentioned above. Any reconciliation will leave some managers unhappy: salesmen will favor goal (1), and production managers will favor goals (2)–(3), perhaps giving different weights to these. Nor is it feasible to produce a global probability function for a system with so many interacting variables.

Example 11. Ordering dessert.

Let's return the restaurant of Example 1. The main course is over, and our agent is offered a dessert menu and the choice of whether to order dessert. On the one hand, there is a direct desire for dessert, perhaps even a craving. This alternative is colored with and motivated by emotion, even if the emotion is not overwhelming. But suppose that there is a contrary emotion. The agent is unhappy with being overweight and has determined to eat less, and may have told others at the table about the decision to undertake a diet. This creates a conflict, coloring the choice of dessert with negative associations, perhaps even shame. The chief difference between this conflict and those in Examples 2 and 4 is that this decision is emotionally "warm;" the outcome may be influenced by a craving and the presence of the desired object. (Perhaps this is why some restaurants invest in dessert trays.)

Example 12. An unanticipated elevator.

A man decides to visit his stockbroker in person, something he has never done. He takes a bus to a stop near the stockbroker's downtown address, gets off the bus, locates the building and enters it. He finds a bank of elevators, and sees that the stockbroker is on the 22nd floor. This man has a strong dislike for elevators, and is not feeling particularly energetic that day. He reconsiders his plan.

Example 13. A woman is working in her garden.

She becomes hot and tired, and decides to take a break. Or she hears the telephone ringing in her house, and decides to answer it. Or she sees smoke coming out of the window of her house, and runs for help.

Example 14. The wrath of Achilles.

In Book I of *The Iliad*, the hero Achilles is outraged and dishonored by his warlord Agamemnon, who insults him and declares that he will take back, in compensation for his own loss and Achilles' disrespectful behavior, the captive woman that Achilles had received as his war prize.

Homer goes on to describe Achilles' reaction. Achilles is headstrong, but his reaction is partly physical and partly intellectual: his heart pounds with rage, but instead of acting immediately he asks himself a question: should he draw his sword and kill the king? To explain his decision, the poet brings in a god: Athena, invisible to everyone else, seizes him by the hair and persuades him to give in and be patient.

For our purposes, we can suppose that Athena is a literary device. The outrage leads to a direct desire to kill, but instead of acting on it, Achilles realizes that it would be better to restrain himself.

Even though it is "hot" — strongly informed by emotion — reasoning intervenes here between the emotional shading of the alternatives and an ensuing resolution to act.

Example 15. Deciding what to say at a given point in a conversation.

Conversation provides many good examples of deliberative reasoning. Where there is conscious deliberation, it is likely to be devoted to content selection. But the reasoning that goes into deciding how to express a given content can be quite complex.

Certainly, any adequate theory of practical reasoning must at least be compatible with this broad range of cases. Better, it should be capable of saying something about the reasoning involved in all of them. Even better, there should be a single architecture for practical reasoning, capable of dealing with the entire range of reasoning phenomena.[3] No doubt, there are special-purpose cognitive modules (e.g., for managing perception, motor behavior, and some aspects of language). But in the absence of convincing, independent psychological evidence it would be perverse to formulate a theory of a special type of practical reasoning, such as preference generation, probability estimation, or means-end reasoning, and to postulate a "cognitive module" that performs just this reasoning. All these types of reasoning can be involved in the same practical problem situation, and interact strongly. This methodology would be likely to produce an *ad hoc* and piecemeal account of practical reasoning.

1.2 Towards a classification

The examples in the previous section suggest a set of features that can be used to classify specimens of deliberative reasoning.

1. Are only a few variables (e.g., desiderata, causal factors, initial conditions) involved in the decision?
2. Do conflicting preferences need to be resolved in making the decision?
3. Is the time available for deliberation small compared to the time needed for adequate reflection?
4. Is the deliberation immediate? That is, will the intentions that result from the deliberation be carried out immediately, or postponed for future execution?
5. Is the deliberation carried out in "real time" as part of an ongoing activity involving sensory and motor activities?
6. Does the reasoning have to interface closely with sensory and motor systems?
7. Is the activity part of a group or team?
8. Does the context provide a definite, relatively small set of actions, or is the set of actions open-ended?
9. Is there certainty about the objective factors that bear on the decision?
10. Is the associated risk small or great?
11. Is the goal of deliberation a single action, or a sequence of actions?
12. Is continuous time involved?
13. Is the deliberation colored with emotions?
14. Is the action habitual, or automatic and unreflective?
15. Is there conscious deliberation?

[3] For the idea of a cognitive architecture, see [38].

16. Are there existing plans in play to which the agent is committed or that already are in execution?

Many of the differences marked by these features are matters of degree, so that the boundaries between the types of reasoning that they demarcate are fluid. This strengthens the case for a general approach to the reasoning. There is nothing wrong with concentrating on a special case to see what can be learned from it. Chess and decision problems that, like Savage's omelet, involve a solution to the "small worlds problem"[4] provide good examples of cases where this methodology has paid off. But to concentrate on these cases without paying any attention to the broad spectrum of examples runs the risk of producing a theory that will not be contribute usefully to something more general.

1.3 Disciplines and approaches

Many different disciplines have something to say about practical reasoning. The main theoretical approaches belong to one of the five following areas.

1. Philosophy
2. Logic
3. Psychology
4. Decision Theory and Game Theory
5. Artificial Intelligence

Of course, there is a good deal of overlap and mixing of these approaches: AI, for instance, is especially eclectic and has borrowed heavily from each of the other fields. But work in each area is colored by the typical problems and methods of the discipline, and — typically, at least — has a distinctive perspective that is inherited from the parent discipline.

The following discussion of these five approaches is primarily interested in what each has to contribute to the prospects for formalizing practical reasoning.

1.3.1 Philosophy

The topic of practical reasoning goes back to Aristotle. In the Twentieth Century there was a brief revival of philosophical interest in the topic of "practical inference." This coincided more or less with early work on deontic and imperative logic, and

[4] This is the problem of framing a decision problem, concentrating only on the factors that are relevant.

was carried out by a group of logically minded philosophers and a smaller group of philosophicaly minded logicians. It is a little difficult to distinguish philosophy from logic in this work; I will more or less arbitrarily classify Kenny and some others as philosophers for the purposes of this exposition, and von Wright as a logician.

Post-Fregean interest in imperative logic seems to have begun about the time of World War 2, with [28, 26, 41]. Later, in the 1960s,[5] some British philosophers became interested in the topic. This period saw 10 or more articles relevant appearing in journals like *Analysis.* Of these, [31] seems to have the most interesting things to say about the problem of formalizing practical reasoning.[6]

Kenny begins with Aristotle's practical syllogism, taking several specimens of means-end reasoning from the Aristotelian corpus, and beginning with the following example, based on a passage in *Metaphysics* 1032b19.

Example 16. A doctor prescribing.

This man is to be healed.
If his humors are balanced, he will be healed.
If he is heated, his humors will be balanced.
If he is rubbed, he will be heated.
So I'll rub him.

The premisses of the reasoning, according to Kenny, are either (i) desires or duties, or (ii) relevant facts. And he characterizes the conclusion as an action.[7] Kenny points out that this sort of reasoning doesn't fit Aristotelian syllogistic, and that a straightforward modern formalization of it would be invalid. To put it crudely, the inference from P, $Q \rightarrow P$, $R \rightarrow Q$, and $S \rightarrow R$ to S is invalid.

Here, I think Kenny has indicated an important type of practical reasoning, and pointed out a glaring problem with the propositional calculus as a formalization medium. Unfortunately, the theory that he proposes in this paper doesn't seem to solve the problem of providing an account of validity that matches the reasoning. In fact, there are many glaring problems with the crude Propositional Calculus formalization of Example 16, involving the deductive formulation of the reasoning as well as the faithfulness of the formalization to the language of the example.

[5] Judging from internal evidence, the work of Richard Hare influenced this episode of interest in the topic. Elizabeth Anscombe [2] may also have been an influence, as well as G.H. von Wright.

[6] For more about this period, see [23].

[7] The Aristotelian texts make it pretty clear that Aristotle considered the conclusion to be an action. But for our purposes, it would work better to think of the conclusion as an expression of intention. In some circumstances — when the deliberation is concerned with immediate action and the reasoning is sufficiently persuasive, there is no gap between intention and action.

The failure of Kenny's proposal and of similar ones at the time seems to originate in a lack of logical resources that do justice to the problem. The Propositional Calculus is certainly not the right tool, and deduction is certainly not the right characterization of the reasoning. The only idea that was explored at the time was that of providing a logic of "imperative inference." This idea might help with one problem: formalizing the first premiss of Example 16, which does not seem like a straightforward declarative. But it can't begin to address the challenge posed by the invalidity of the argument. Besides, the idea of an imperative logic didn't lead to anything very new, because of another trend that was taking place at about the same time.

This trend, which tried to absorb imperative and practical inference into some sort of modal logic, was also underway in the 1960s. [32] provides a logic of imperatives that prefigures the Stit approach of [6], hence a modal approach that brings in the idea of causing a state of affairs. And [11], recommends and develops a reduction of imperative logic to a more standard deontic logic. This idea provides formal systems with excellent logical properties. But it does so at the expense of changing the subject and leaving the central problem unsolved. Reasoning in deontic logic is deductive, and if you formalize typical specimens of means-end reasoning like Example 16 in these systems, the formalizations will be invalid.

Even though the literature shows a sustained series of attempts in this period to formalize practical inference, the work didn't lead to anything like a consensus, and produced no sustainable line of logical development. In retrospect, we can identify several assumptions that rendered the formalization project unsustainable:

1. These philosophers relied too much on deductive inference, with the propositional calculus as a paradigm, and too little on models;
2. They tended to work with overly simple formal languages;
3. They didn't bring actions into the formalization explicitly;
4. They missed the insight that means-end reasoning is more like abduction or heuristic search than deduction.

As we will see in Section 1.3.5, more recent and quite separate developments in computer science have yielded sophisticated logics of means-end reasoning, effectively solving the formalization problem that led to an earlier philosophical impasse in the 1960's and 1970's. The moral seems to be that formalization projects of this sort can involve multiple challenges, and that it can be hard to address these challenges without a body of applications and a community of logicians committed to formalizing the applications and mechanizing the reasoning.

Meanwhile, philosophers seem to have drawn the conclusion that close attention to the reasoning, and searching for formalizations, is not likely to be productive. In the more recent philosophical work on practical reasoning, it is actually quite difficult to find anything that bears on the formalization problem. Almost entirely, the philosophical literature is devoted to topics that might serve to provide philosophical foundations for the theory of practical reasoning — if there were such a theory. Even if, as Elijah Millgram claims in [36], the driving issue in the philosophy of practical reasoning is to determine which forms of practical reasoning are correct, philosophers seem to pursue this inquiry with informal and very loose ideas of the reasoning itself. In many cases — for instance, the issue of whether intentions cause actions — no formalization of the reasoning is needed for the philosophical purposes. In other cases, however, a formal theory of practical reasoning might help the philosophy, refining some old issues and suggesting new ones.

Even though some philosophers maintain positions that would sharply limit the scope of practical reasoning (reducing it, for instance, to means-end reasoning), I don't know of any explicit, sustained attempt in the philosophical literature to delineate what the scope of practical reasoning should be. I don't see how to do this without considering a broad range of examples, as I try to do above in Section 1.1. But in fact, examples of practical reasoning are thin on the ground in the philosophical literature; in [36], for instance, I counted only 12 examples of practical reasoning in 479 pages — and many of these were brief illustrations of general points.

1.3.2 Logic

There are few departments of logic, and work in logic bearing on practical reasoning tends to be carried out in the context of either Philosophy or Computer Science, and to be influenced by the interests of the parent disciplines. There are, in fact, two separate strands of logical research, one associated with Philosophy and the other with Artificial Intelligence. These have interacted less than one might wish.

Philosophical logic. Georg Henrik von Wright was explicitly interested in practical reasoning, from both a philosophical and a logical standpoint. Most of his writings on the topic are collected together in [56]; these were published between 1963 and 1982. Like Kenny, von Wright begins with Aristotle's practical syllogism. But he avoids the problem of invalidity by strengthening premisses that introduce ways of achieving something. Von Wright's version of Example 16 would look like this:

I want to heal this man.

> Unless his humors are balanced, he will not be healed.
> Unless he is heated, his humors will not be balanced.
> Unless he is rubbed, he will not be heated.
> Therefore I must rub him.

By departing from Aristotle's formulation, von Wright makes it easier to formulate the inference in a deontic logic, and to see how the formalization might be valid. At the same time, he is making it more difficult to fit the formalization to naturally occurring reasoning. As in this example, where, for instance, there is surely more than one way to heat the patient, the means that a deliberator chooses in typical means-end reasoning will not be the only way to achieve the end.

This simplification makes it easier for von Wright to propose modal logic, and in particular deontic logic, as the formalization medium for practical reasoning. Von Wright also characterizes his version of deontic logic as a "logic of action," but all this seems to mean is that the atomic formulas of his language may formalize things of the form 'Agent A does action a.' He has little or nothing to say about reasoning about action.

I will not say much here about the subsequent history of deontic logic as a part of philosophical logic. As the field developed, it acquired its own problems and issues (such as the problem of reparational obligations), but as philosophers concentrated on declarative formalisms and deductive logic, the relevance to practical reasoning, and even means-end reasoning, that von Wright saw in his in early papers such as [55], attenuated.

Although the subsequent history of deontic logic was less directly concerned with practical reasoning, it shows a healthy tendency to concentrate on naturally occurring problems that arise in reasoning about obligation. This work has a place in any general theory of practical reasoning. Obligations play a role as constraints on means-end reasoning, and reasoning about obligations has to be flexible to cope with changing circumstances.

Also, the problem of modeling conditional obligations has produced a large literature on the relationship between modal logic and preference.[8] Of course, reasoning about preferences intrudes into practical reasoning in many ways. How to fit it in is something I am not very clear about at the moment; part of the problem is that so many different fields study preferences, and preferences crop in so many different types of practical reasoning. Maybe the best thing would be to incorporate preferences in a piecemeal way, and hope that a more general and coherent approach might emerge from the pieces.

[8] See, for instance, [24, 27].

The Stit approach to agency was already mentioned in Section 1.3.1. This provides a model-theoretic account of how actions are related to consequences that is quite different from the ones that emerged from the attempts in AI to formalize planning. The connections of Stit theory to practical reasoning are tenuous, and I will not have much to say about it.

Philosophy and philosophical logic have served over the years as a source of ideas for extending the applications of logic and developing logics that are appropriate for the extensions. One would hope that philosophy would continue to play this role. But — at least, for areas of logic bearing on practical reasoning — the momentum has shifted to computer science, and especially to logicist AI and knowledge representation. This trend began around 1980, and has accelerated since. Because many talented logicians were attracted to computer science, and because the need to relate theories to working implementations provided motivation and guidance of a new kind, this change of venue was accompanied by dramatic logical developments, and improved insights into how logic fits into the broader picture. I would very much like to see philosophy continue to play its foundational and creative role in developing new applications of logic, but I don't see how this can happen in the area of practical reasoning unless philosophers study and assimilate the recent contributions of computer scientists.

The point is illustrated by [19]. The paper is rare among contemporary papers in urging the potential importance of a logic of practical reasoning, but — in over 100 pages — it is unable to say what a coherent, sustained research program on the topic might be like. It does mention some important ideas, such as taking the agent into account, as well as nonmonotonic and abductive reasoning, but offers no explicit, articulated theories and in fact is hesitant as to whether logic has a useful role to play, repeating some doubts on this point that have been expressed by some roboticists and cognitive psychologists. Although it cites a few papers from the AI literature, the citations are incidental; work on agent architectures, abductive reasoning, and means-end reasoning goes unnoticed. Part of the problem is that the authors seem to feel that work in "informal logic" might be useful in approaching the problem of practical reasoning — but the ideas of informal logic are too weak to provide any helpful guidance. If we are interested in accounting for the practical reasoning of agents, we have to include computer programs. For this, we need formal logic — but formal logic that is applicable.

I couldn't agree more with Gabbay and Woods that logicians should be concerned with practical reasoning. But to make progress in this area, we need to build on the accomplishments of the formal AI community.

1.3.3 Psychology

From the beginning of cognitive psychology, a great deal of labor has gone into collecting protocols from subjects directly engaged in problem-solving, much of it practical. Herbert Simon and Allen Newell were early and peristent practitioners of this methodology. This material contains many useful examples; in fact, it helped to inspire early characterizations of means-end reasoning in Artificial Intelligence.

As early as 1947, in [47], Simon had noted divergences between decision-making in organizations and the demands of ideal rationality that are incorporated in decision theory; he elaborated the point in later work. An important later trend that began in psychology, with the work of Amos Tversky and Daniel Kahneman, studies these differences in more detail, providing many generalizations about the way people in fact make decisions and some theoretical models; see, for instance, [29, 53].

Tversky and Kahneman's experimental results turned up divergences between ideal and actual choice-making that were not obviously due, as Simon had suggested, merely to the application of limited cognitive resources to complex, time-constrained problems. Since their pioneering work, this has become a theme in later research.

All this raises a challenging foundational problem, one that philosophers might be able to help with, if they gave it serious attention. What level of idealization is appropriate in a theory of deliberation? What is the role of "rationality" in this sort of idealization? Is there a unique sort of rationality for all practically deliberating agents, or are there many equally reasonable ways of deliberation, depending on the cognitive organization and deliberative style of the agent? Is the notion of rationality of any use at all, outside the range of a very limited and highly idealized set of decision problems? Probably it would be unwise to address these problems before attempting to provide a more adequate formalization of practical reasoning — that would be likely to delay work on the formalization indefinitely. But the problems are there.

Nowadays, the cognitive psychology of decision-making has migrated into Economics and Management Science, and is more likely to be found in economics departments and schools of business than in psychology departments. This doesn't affect the research methods much, but it does improve the lines of communication between researchers in behavioral economics and core areas of economics. As a result, economic theorists are becoming more willing to entertain alternatives to the traditional theories.

1.3.4 Decision Theory and Game Theory

The literature in these areas, of course is enormous, and most of it has to do with practical reasoning. But traditional work in game theory and decision theory concentrates on problems that can be formulated in an idealized form — a form in which the reasoning can be reduced to deriving an optimum result by calculation.[9] As a result, work in this tradition tends to neglect much of the reasoning in practical reasoning. Of course, an agent must reason to wrestle a practical problem into the required form — to solve Savage's "small worlds problem" — but the literature in economics tends to assume that somehow the problem has been framed, without saying much if anything about the reasoning that might have gone into this process. (Work in decision analysis, of course, is the exception.) And once a problem has been stated in a form that can be solved by calculation, there is little point in talking about deliberative processes.

If we are concerned with the entire range of examples presented in Section 1.1, however, we find many naturally occurring problems that don't fit this pattern; and some of these, at least, exhibit discursive, inferential reasoning. This is one reason why I believe that a general theory of practical reasoning will reserve an important place for qualitative reasoning, and especially for inferential reasoning — the sort of reasoning that gives formalization and logic a foothold. In this respect, Aristotle was on the right track.

At the very least, practical reasoning can involve inference and heuristic search, as well as calculation. (Calculation, of course, is a form of reasoning, but is not inferential, in the sense that I intend.) Any theory of practical reasoning that emphasizes one sort of reasoning at the expense of others must sacrifice generality, confining itself to only a small part of the territory that needs to be covered by an adequate approach. The imperialism of some of those (mainly philosophers, these days) who believe that there is nothing to rationality or practical reasoning other than calculations involving probability and utility, can partly be excused by the scarcity of theoretical alternatives. I will argue in this article that the field of Artificial Intelligence has provided the materials for developing such alternatives.

As I said in Section 1.3.3, research in behavioral economics has made microe-

[9] Microeconomists and statisticians are not the only ones who have taken this quantitative, calculational paradigm to heart. Many philosophers have accepted the paradigm as a model of practical reasoning and rationality. See, for instance, [49], a book-length study of practical deliberation, which takes the only relevant theoretical paradigms to be decision theory and game theory, and takes them pretty much in the classical form. Skyrms' book and the many other philosophical studies along these lines have useful things to say; my only problem with this literature is the pervasive assumption that practical reasoning can be comprehensively explained by quantitative theories based on the assumption that agents have global probability and utility functions.

conomists generally aware that, in their original and extreme form, the idealizations of decision theory don't account well for a broad range of naturally occuring instances of practical reasoning. Attempts to mechanize decision-making led computer scientists to much the same conclusion.

A natural way to address this problem begins with decision theory in its classical form and attempts to relax the idealizations. Simon made some early suggestions along these lines; other, quite different proposals can be found in [57] and [42]. And other relaxations of decision theory have emerged in Artificial Intelligence: see the discussion of Conditional Preference Nets below, in Section 1.3.5. Still other relaxations have emerged out of behavioral economics, such as Tversky and Kahneman's Prospect Theory; see [29].

Programs of this sort are perfectly compatible with what I will propose here. A general account of practical reasoning has to include calculations that somehow combine probability (represented somehow) and utility (represented somehow), in order to estimate risk. The more adaptable these methods of calculation are to a broad range of realistic cases, the better. I do want to insist, however, that projects along these lines can only be part of the story. Anyone who has monitored their own decision making must be aware that not all practical reasoning is a matter of numerical calculation; some of it is discursive and inferential. A theory that does justice to practical deliberation has to include both forms of reasoning. From this point of view, the trends from within economics that aim at practicalizing game theory and decision theory are good news. From another direction, work in Artificial Intelligence that seeks to incorporate decision theory and game theory into means-end reasoning is equally good news.[10]

In many cases of practical reasoning, conflicts need to be identified and removed or resolved. Work by economists on value tradeoffs is relevant and useful here; the classical reference is [30], which contains analyses of many naturally occuring examples.

1.3.5 Computer science and artificial intelligence

For most of its existence, the field of AI has been concerned with realistic decision problems, and compelled to formalize them. As the field matured, the AI community looked beyond procedural formalizations in the form of programs to declarative formulations and logical theories. Often AI researchers have had to create their

[10] For a survey, now getting rather old, see [7]. For an example of a more recent, more technical paper, see [43].

own logics for this purpose.[11] Here, I will be concerned with three trends in this work: those that I think can offer most to the formalization of practical reasoning. These three are: means-end reasoning, reasoning about preferences, and agent architectures.

Dynamic logic and imperative inference. When an agent is given instructions and intends to carry them out unquestioningly, there is still reasoning to be done, and the reasoning is practical[12] — although, as the instructions become more explicit, the less scope there is for interesting reasoning from the human standpoint. Even so, the case of computer programs, where explicitness has to be carried out ruthlessly, can be instructive, because it shows how logical theory can be useful, even when the reasoning paradigm is not deductive.

A computer program is a (possibly very large and complex) imperative, a detailed instruction for carrying out a task. Many of its components, such as

$$\textbf{let } y \textbf{ be } x$$

("set the value of y to the current value of x") are imperatives, although some components, like the antecedent of the conditional instruction

$$\textbf{if } (x < y \textbf{ and } \textbf{not}(x = 0)) \textbf{ then let } z \textbf{ be } y/x$$

are declarative.

Inference, in the form of proofs or a model theoretic logical consequence relation, plays a small part in the theory of dynamic logic. Instead, *execution* is crucial. This idea is realized as the series of states that the agent (an idealized computer) goes though when, starting in a given initial state, it executes a program. Because states can be identified with assignments to variables, there are close connections to the familiar semantics of first-order logic.

Dynamic logic is useful because of its connection to *program verification.* A program specification is a condition on what state the agent will reach if it executes the program; if the initial state of a parsing program for an English grammar G, for instance, describes a string of English words, the program execution should eventually halt. Furthermore, (1) if the string is grammatical according to G, the executor should reach a final state that describes a parse of the string, and (2) if the

[11] Throughout his career, John McCarthy was a strong advocate of this approach, and has done much of the most important work himself. See [35] for an early statement of the methodology, and a highly influential proposal about how to formalize means-end reasoning.

[12] See [33].

string is not grammatical according to G it should reach a final state that records its ungrammaticality.

Dynamic logic has led to useful applications and has made important and influential contributions to logical theory. It is instructive to compare this to the relatively sterile philosophical debate concerning "imperative inference" that took place in the 1960s and early 1970s.[13] To a certain extent, the interests of the philosophers who debated imperative inference and the logicians who developed dynamic logic were different. Among other things, the philosophers were interested in applications to metaethics, and computational applications and examples didn't occur to them.

But the differences between philosophers and theoretical computer scientists, I think are relatively unimportant; some of the philosophers involved in the earlier debate were good logicians, and would have recognized a worthwhile logical project if it had occurred to them. In retrospect, three factors seem to have rendered the earlier debate unproductive:

1. Too great a reliance on deductive paradigms of reasoning;
2. Leaving a model of the executing agent out of the theoretical picture;
3. Confining attention to simple examples.

In dynamic logic, the crucial semantic notion is the correctness of an imperative with respect to a specification. Logically interesting examples of correctness are not likely to present themselves without a formalized language that allows complex imperatives to be constructed, and without examples of imperatives that are more complicated than 'Close the door'. (The first example that is presented in [25] is a program for computing the greatest common divisor of two integers; the program uses a *while*-loop.) And, of course, a model of the executing agent is essential to the logical theory. In fact, what is surprising is how much logic can be accomplished with such a simple and logically conservative agent model.

As I said, the activity of interpreting and slavishly executing totally explicit instructions is a pretty trivial form of practical reasoning. But a logic of this activity is at least a start. I want to suggest that, in seeking to formalize practical reasoning, we should be mindful of these reasons for the success of dynamic logic, seeking to preserve and develop them as we investigate more complex forms of practical reasoning.

Planning and the formalization of means-end reasoning. Perhaps the most important contribution of AI to practical reasoning is the formalization of means-

[13] See, for instance, [58, 20] as well as [31], which was discussed above, in Section 1.3.1.

end reasoning, along with appropriate logics, and an impressive body of research into the metamathematical properties of these logics, and their implementation in planning systems.[14]

This approach to means-ends reasoning sees a planning problem as consisting of the following components:

1. An initial state. (This might be described by a set of literals, or positive and negative atomic formulas.)
2. Desiderata or goals. (These might consist of a set of formulas with one free variable; a state that satisfies these formulas is a goal state.)
3. A set of actions or operators. Each action a is associated with a causal axiom, saying that if a state s satisfies certain preconditions, then a state $\textsc{Result}(a, s)$ that results from performing a in s will satisfy certain postconditions.

Here, the fundamental logical problem is how to define the state or set of successor states[15] resulting from the performance of an action in a state. (Clearly, not all states satisfying the postconditions of the action will qualify, since many truths will carry over to the result by "causal inertia.") This large and challenging problem spawned a number of subproblems, of which the best-known (and most widely misunderstood) is the *frame problem.* Although no single theory has emerged from years of work on this problem as a clear winner, the ones that have survived are highly sophisticated formalisms that not only give intuitively correct results over a wide range of test cases, but provide useful insights into reasoning about actions. Especially when generalized to take into account more realistic circumstances, such as uncertainty about the current state and concurrency or nondeterminism, these planning formalisms deliver logical treatments of means-end reasoning that go quite far towards solving the formalization problem for this part of practical reasoning.

I will try to say more about how these developments might contribute to the general problem of formalizing practical reasoning below, in Section 2.3.

Reasoning about preferences. It is hard to find AI applications that don't involve making choices. In many cases, it's important to align these choices with the designer's or a user's preferences. Implementing such preference-informed choices requires (i) a representation framework for preferences, (ii) an elicitation method that yields a rich enough body of preferences to guide the choices that need to be made, and (iii) a way of incorporating the preferences into the original algorithm.

[14] [1] is a collection of early papers in the field. Both [46] and [39] describe the earlier logical frameworks and their later generalizations; [39] also discusses implementation issues.

[15] Depending on whether we are working with the deterministic or the nondeterministic case.

Any attempt to extract the utilities needed for even a moderately complex, realistic decision problem will provide motives for relaxing the classical economic models of utility; but the need for workable algorithms seems to sharpen these motives. See [22] for examples and details, and [14], which provides a wide-ranging foundational discussion of the issues, with many references to the economics literature.

Of the relaxations of preference that have emerged in AI, *Ceteris Paribus* Preference Nets are one of the most widely used formalisms.[16] As in multi-attribute utility theory, the outcomes to be evaluated are characterized by a set of features. A parent-child relation must be elicited from a human subject; this produces a graph called a *CP-net*. The parents of a child feature are the features that directly influence preferences about the child. For instance, the price of wheat in the fall (high or low) might influence a farmer's preferences about whether to plant wheat in the spring. If the price will be high, the farmer prefers to plant wheat; otherwise, he prefers not to plant it. On the other hand, suppose that in the farmer's CP-net the price of lumber is unrelated to planting wheat. It can then be assumed that preferences about planting wheat are independent of the price of lumber.

To complete the CP-net, a preference ranking over the values of a child feature must be elicited for each assignment of values to each of the parent features.

Acyclic CP-nets support a variety of reasoning applications (including optimization), and — combined with means-end reasoning — provide an approach to preference-based planning.[17] And in many realistic cases it is possible to extract the information needed to construct a CP-net.

There are extensions of this formalism that allow for a limited amount of reasoning about the priorities of features in determining overall preferences; see [9].

The work in AI on preferences, like decision analysis, tends to concentrate on extracting preferences from a user or customer. Practical reasoning, however, produces a different emphasis. Some of the examples in Section 1.1 — for instance, Examples 1, 4, 10, and 11 — were designed to show that preferences are not automatically produced by the environment, by other agents, by the emotions, or by a combination of these things. We deliberate about what is better than what, and preferences can be the outcome of practical reasoning.[18] The status of an agent trying to work out its own preferences, and of a systems designer or decision analyst trying to work out the preferences of a person or an organization, may be similar in some ways, but I don't think we can hope that they are entirely the same. Nevertheless, insights into methods for extracting preferences from others might be helpful in thinking about how we extract our own preferences.

[16] See, for instance, [13, 8].

[17] See [4] for details and further references.

[18] For some preliminary and sketchy thoughts about this, see [52].

Agent architectures. A nonexecuting planning agent is given high-level goals by a user, as well as the declarative information about actions and the current state of things, as well perhaps as preferences to be applied to the planning process. With this information, it performs means-end reasoning and passes the result along to the user in the form of a plan.

This agent is not so different from the simple instruction-following agent postulated by dynamic logic; its capabilities are limited to the execution of a planning program, and it has little or no autonomy. But — especially in time-limited planning tasks — it may be difficult to formulate a specification, because the notion of what counts as an optimal plan in these condition is unclear.

When the planning agent is equipped with means of gathering its own information, perhaps by means of sensors, and is capable of performing its own actions, the situation is more complicated, and more interesting. Now the agent is interacting directly with its environment, and not only produces a plan, but must adopt it and put it in to action. This has a number of important consequences. The agent will need to perform a variety of cognitive functions, and to interleave cognitive performances with actions and experiences.

1. Many of the agent's original goals may be conditional, and these goals may be activated by new information received from sensors. This is not full autonomy, but it does provide for new goals that do not come from a second party.
2. Some of these new goals may be urgent; so the agent will need to be interruptable.
3. It must commit to plans — that is, it must form intentions. These intentions will constrain subsequent means-end reasoning, since conflicts between its intentions and new plans will need to be identified and eliminated.
4. It will need to schedule the plans to which it has committed.
5. It will need to monitor the execution of its plans, to identify flaws and obstacles, and repair them.

Recognizing such needs, some members of the AI community turned their attention from inactive planners to *agent architectures*, capable of integrating some of these functions. Early and influential work on agent architectures was presented in [10]; this work stressed the importance of intentions, and the role that they play in constraining future planning.

Any means-end reasoner needs desires (in the form of goals) and beliefs (about the state of the world and the consequences of actions). As Bratman, Israel, and Pollock point out, an agent that is implementing its own plans also needs to have intentions. Because of the importance of these three attitudes in the work that was influenced by these ideas, architectures of this sort are often known as *BDI architectures.* For an extended discussion of BDI architectures, with references to the literature up to 2000, see [59]. See also [21].

Work in "cognitive robotics" provides a closely related, but somewhat different approach to agent architectures. Ray Reiter, a leading figure in this area, developed methods for integrating logical analysis with a high-level programming language called GoLog, an extension of Prolog. Reiter's work is continued by the Cognitive Robotics Group at the University of Toronto.

Developments in philosophical logic and formal semantics have provided logics and models for propositional attitudes; for instance, see [15, 16]. Using these techniques, it is possible to formulate a metatheory for BDI agency. Such a metatheory is not the architecture; the reasoning modules of a BDI agent and overall control of reasoning has to be described procedurally. But the metatheory can provide specifications for some of the important reasoning tasks. Wooldridge's logic of rational agents, $\mathcal{LORA}$, develops this idea; see [59].

A final word. Logicist AI has struggled to maintain a useful relation to applications, in the form of workable technology. Although the struggle has been difficult, many impressive success stories have emerged from this work — enough to convince the larger AI community of the potential value of this approach. The incentive to develop working applications has, I believe, been very helpful for logic, enabling new ideas that would not have been possible without the challenges posed by complex, realistic reasoning tasks.

Practical reasoning is not quite the same as logicist AI, or even the logical theory of BDI agents. But the successful use of logical techniques in this area of AI provides encouragement for a logical approach to practical reasoning. And, of course, it provides a model for how to proceed.

2 Towards a formalization

The challenge is this: how to bring logical techniques to bear on practical reasoning, and how to do this in a way that is illuminating, explanatory, and useful? In the rest of this article, I will only try to provide an agenda for addressing this challenge. The agenda divides naturally into subprojects. Some of these subprojects can draw on existing work, and especially on work in AI, and we can think of them as well underway or even almost completed. Others are hardly begun.

2.1 Relaxing the demands of formalization

Let's return to the division between theoretical and practical reasoning.

Traditionally, domains that involve theoretical reasoning are formalized using what Alonzo Church called the "logistic method."[19] This method aims to formulate a formal language with an explicit syntax, a model-theoretically characterized consequence relation, and perhaps a proof procedure. Traditional formalizations did not include a model of the reasoning agent, except perhaps, in the highly abstract form of a Turing machine — this sort of agent is guaranteed whenever the consequence relation is recursively enumerable.

When it comes to practical reasoning, I believe that we have to be prepared to relax Church's picture of logical method.[20]

My own proposal for a relaxation is this: (1) we need to add a model of the reasoning agent, (2) we need to identify different phases of practical reasoning in agent deliberation, and different ways in which logic might be involved in each phase of the reasoning, and (3) consequently, we need to be prepared to have a logical treatment that is more pluralistic and less unified.

2.2 Agent architectures and division of logical labor

How should we model an agent that is faced with practical reasoning problems? In Section 1.1, I suggested that we should aim at, or at least acknowledge the existence of, a very broad range of reasoning problems. Suppose, for instance, that we classify the types of reasoning that we may need to consider in terms of the sort of conclusion that is reached. In view of the examples that were presented in Section 1.1, we will need to be prepared for the agent to infer:

1. Goals, which then invoke planning processes;
2. Plans, and the subgoals or means that emerge from plans;
3. Preferences emerging from reasoning about tradeoffs and risk;
4. Intentions, commitments about what to do, and (to an extent) about when to do it;
5. Immediate decisions about what plan to execute;

[19] [12][pp. 47–58].

[20] In fact, writing in 1956, Church was uncomfortable with semantics and model theory. He included these topics, but in a whisper, using small type. Over 50 years later, we have become quite comfortable with model theory and semantics, and are more likely to insist on this ingredient than on proof procedures. And in areas where logic is applied, we have become increasingly willing to bring the reasoning agent into the picture.

6. Immediate, engaged adjustments of ongoing activities and plan executions, and shifts of attention that can affect the task at hand.

The examples in Section 1.1 were chosen, in part, to illustrate these activities. These sorts of deliberation are distinct, and all are practical. Although some of them can be automatic, they all can involve deliberate reasoning.

These six activities comprise my (provisional) division of practical reasoning into subtasks, and of the deliberating agent into subsystems. Each of them provides opportunities for logical analysis and formalization. I will discuss them in turn.

2.3 Means-end reasoning

This is the best developed of the six areas. We can refer to the extensive AI literature on planning and means-end reasoning not only for well developed logical theories, but for ideas about how this deliberative function interacts with the products of other deliberative subsystems — for instance, with preferences, and with plan monitoring and execution.

2.4 The practicalization of desires

On the other hand, work in AI on means-end reasoning, and on BDI agents, has little or nothing to say about the emotions and the origins of desires. In general, it is assumed that these come from a user — although the goals may be conditional, so that they are only activated in the appropriate circumstances. In principle, there is no reason why goals couldn't be inferred or learned. But the relevant reasoning processes have not, as far as I know, been formalized.

In truly autonomous agents some desires — perhaps all — originate in the emotions. Although a great deal has been written about the emotions, it is hard to find work that could serve a useful logical purpose.[21]

However desires originate, although they may be emotionally colored, they may not all be emotionally "hot." And to be useful in reasoning, some desires must be conditional, and self-knowledge about conditional desires must be robust. My preference for white wine this evening will probably be accompanied by feelings of pleasure when I think about the refreshing taste of white wine. But the feeling of hypothetical pleasure is relatively mild; I am certainly not carried away by the

[21] Not [50], which has a chapter on "Reason and the passions," a section on "The Rationality of the emotions," and a chapter on "The logic of the emotions." Not [37], written by an author who knows something about AI. But work on modeling artificial characters for applications in areas like interactive fiction might be useful; see [5].

feeling. AI systems builders are interested in obtaining a large body of conditional preferences from users because preferences need to be brought to bear under many different circumstances. Therefore a user's unconditional preferences — the preferences that are activated in the actual state of affairs — will not in themselves be very useful. Fully autonomous agents need conditional preferences as well, in planning future actions and in contingency planning.

Perhaps — to develop the example of preference for white wine a bit further — the only mechanism that is needed to generate conditional desires is the ability to imagine different circumstances, together with the ability to color these circumstances as pleasant (to some degree), and unpleasant (to some degree). But it is unlikely to be this simple, because pleasantness is not monotonic with respect to information: I find the idea of a glass of white wine quite pleasant, but the idea of a glass of white wine with a dead fly in it quite unpleasant. Also, my feelings about some imagined situations can be mixed, with elements that I find pleasant and elements that I find unpleasant. At this point, I might have to invoke a conflict resolution method that has little or nothing to do with the emotions.

This leads to a further point: there is a difference between raw or immediate desires, or *wishes*, and all-things-considered desires, or *wants*. This is because desires can not only conflict with one another, but with beliefs. And, when they conflict with beliefs, desires must be overridden: to do otherwise would be to indulge in wishful thinking.

In [51], I explored the possibility of using a nonmonotonic logic to formalize this sort of practicalization of desires. The target reasoning consisted of deliberations such as the following. (The deliberator is a hiker who forgot her rain gear.)

1. I think it's going to rain.
2. IIf it rains, I'll get wet.
3. If I get wet, I'll stay wet unless I give up and go home.
4. I wouldn't like to stay wet.
5. I wouldn't like to give up and go home.

The argument reaches an impasse, and a conflict needs to be addressed to resolve it. There are two possible conclusions here, depending on how the conflict is resolved:

6. On the whole, I'd rather go home.

6′. On the whole, I'd rather go on hiking.

The main purpose of Steps 1–5 is to identify the conflict.

I'm not altogether happy with the theory presented in [51], but I still believe that the practicalization of desires is an important part of practical reasoning that provides opportunities for using logic to good advantage.

2.5 Intention formation

The product of successful means-end deliberation will be an intention, taking the form of commitment to a plan. But the deliberation would not get started without a goal — and I see no difference between a goal and a provisional and (perhaps very general and sketchy) intention. Often, even in human agents, these goals come from habits, or from compliantly accepted instructions from other agents.

But sometimes goals arise internally, as outcomes of deliberation. The hiker in Section 2.4 provides an example. If the conclusion of the reasoning is a practicalized desire to turn back and head for home, commitment to the conclusion will produce an intention, which may even become a goal for means-end reasoning. ("How am I to get home?")

This is why practicalization can be an important component of practical reasoning, especially if the reasoner is an autonomous human being.

2.6 What to do now?

In the life of an autonomous agent, moments will arise when there is scope for new activities. These opportunities need to be recognized, and an appropriate task needs to be selected for immediate execution. A busy agent with many goals and a history of planning may have a ready-made agenda of tasks for such occasions; but even so, it may take reasoning to select a task that is rewarding and appropriate. I do not know if any useful work has been done on this reasoning problem.

2.7 Scheduling, execution and engagement

Some of the examples in Section 1.1 were intended to illustrate the point that there can be deliberation even in the execution of physically demanding, real-time tasks. And there can be such a thing as overplanning, since the plans that an agent makes and then performs will need to be adjusted to circumstances, and more detailed plans will require more elaborate adjustments.

Also, not all intentions are immediate. Those that are not immediate need to be invoked when the time and occasion are right.

There has been a great deal of useful work on these topic in AI; just one one recent example is [18].

2.8 Framing a practical problem

Leonard Savage's "Small worlds problem" is replicated in the more qualitative setting of means-end deliberation. A means-end reasoning problem requires (at least) a set of actions, a description of the initial conditions, and a goal. But, even in complex cases, formulations of planning problems don't include every action an agent might perform, or every fact about the current state of the world. Somehow, a goal (like "getting to the airport") has to suggest a method of distinguishing the features of states (or "fluents") and the actions that are relevant and appropriate.

I'm sure that ontologies would be helpful in addressing this problem, but other than this I have very little to say about it at the moment.

References

[1] James Allen, James Hendler, and Austin Tate, editors. *Readings in Planning.* Morgan Kaufmann, San Mateo, California, 1990.

[2] G.E.M. Anscombe. *Intention.* Blackwell Publishers, Oxford, 1958.

[3] Minoru Asada, Hiroaki Kitano, Itsuki Noda, and Manuela Veloso. RoboCup today and tomorrow—what we have learned. *Artificial Intelligence*, 110(2):193–214, 1999.

[4] Jorge A. Baier and Sheila A. McIlraith. Planning with preferences. *The AI Magazine*, 29(4):25–36, 2008.

[5] Joseph Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[6] Nuel D. Belnap, Jr., Michael Perloff, and Ming Xu. *Facing the Future: Agents and Choices in Our Indeterminist World.* Oxford University Press, Oxford, 2001.

[7] Jim Blythe. An overview of planning under uncertainty. In Michael Wooldridge and Manuela Veloso, editors, *Artificial Intelligence Today*, pages 85–110. Springer-Verlag, Berlin, 1999.

[8] Craig Boutilier, Ronen I. Brafman, Carmel Domschlak, Holger H. Hoos, and David Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2003.

[9] Ronen I. Brafman, Carmel Domshlak, and Solomon E. Shimony. On graphical modeling of preference and importance. *Journal of Artificial Intelligence Research*, 25:389–424, 2006.

[10] Michael E. Bratman, David Israel, and Martha Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[11] Brian Chellas. *The Logical Form of Imperatives.* Perry Lane Press, Stanford, California, 1969.

[12] Alonzo Church. *Introduction to Mathematical Logic, Vol. 1.* Princeton University Press, Princeton, 1959.

[13] Carmel Domschlak. *Modeling and Reasoning about Inferences with CP-Nets.* Ph.d. dissertation, Ben-Gurion University of the Negev, Be'er Sheva, 2002.

[14] Jon Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111–136, 2004.

[15] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge.* The MIT Press, Cambridge, Massachusetts, 1995.

[16] Melvin Fitting. Intensional logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* The Metaphysics Research Lab, Stanford, California, 2009.

[17] J. Fox and M. Clarke. Towards a formalization of arguments in decision making. In *Proceedings of the 1991 AAAI Spring Symposium on Argument and Belief*, pages 92–99. AAAI, 1991.

[18] Christian Fritz. *Monitoring the Generation and Execution of Optimal Plans.* Ph.d. dissertation, University of Toronto, Toronto, 2009.

[19] Dov M. Gabbay and John Woods. The practical turn in logic. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume XIII*, pages 1–122. Springer-Verlag, Berlin, 2 edition, 2005.

[20] Peter T. Geach. Imperative inference. *Analysis*, 23, Supplement 1(3):37–42, 1963.

[21] Michael Georgeff, Barney Pell, Martha Pollack, Miland Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In Jörg P. Müller, Munidar P. Singh, and Anand S. Rao, editors, *Intelligent Agents V: Agents Theories, Architectures, and Languages*, pages 1–10. Springer-Verlag, Berlin, 1999.

[22] Judy Goldsmith and Ulrich Junker. Preference handling for artificial intelligence. *The AI Magazine*, 29(4):9–12, 2008.

[23] Mitchell Green. The logic of imperatives. In E. Craig, editor, *The Routledge Encyclopedia of Philosophy*, pages 717–21. Routledge, New York, 1997.

[24] Sven Ove Hansson. Preference logic. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume IV*, pages 319–394. Kluwer Academic Publishers, Amsterdam, 2 edition, 2001.

[25] David Harel, Dexter Kozen, and Jerzy Tiuryn. *Dynamic Logic.* The MIT Press, Cambridge, Massachusetts, 2000.

[26] Albert Hofstadter and J.C.C. McKinsey. On the logic of imperatives. *Philosophy of Science*, 6:446–457, 1939.

[27] Andrew J.I. Jones and José Carmo. Deontic logic and contrary-to-duties. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume VIII*, pages 265–344. Kluwer Academic Publishers, Amsterdam, 2 edition, 2002.

[28] Jørgen Jørgensen. Imperatives and logic. *Erkenntnis*, 7:288–296, 1937-1938.

[29] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[30] Ralph H. Keeney and Howard Raiffa. *Decisions With Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley and Sons, Inc., New York, 1976.

[31] Anthony J. Kenny. Practical inference. *Analysis*, 26(3):65–75, 1966.

[32] Edward John Lemmon. Deontic logic and the logic of imperatives. *Logique et Analyse*, 8:39–71, 1965.

[33] David K. Lewis. A problem about permission. In Esa Saarinen, Risto Hilpinen, Ilkka Niiniluoto, and Merrill Province Hintikka, editors, *Essays in Honour of Jaakko Hintikka*. D. Reidel Publishing Co., Dordrecht, Holland, 1979.

[34] Vladimir Lifschitz, Norman McCain, Emilio Remolina, and Armando Tacchella. Getting to the airport: The oldest planning problem in AI. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 147–165. Kluwer Academic Publishers, Dordrecht, 2000.

[35] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Bernard Meltzer and Donald Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[36] Elijah Millgram. Practical reasoning: The current state of play. In Elijah Millgram, editor, *Varieties of Practical Reasoning*, pages 1–26. The MIT Press, Cambridge, Massachusetts, 2001.

[37] Marvin Minsky. *The Emotion Machine.* Simon & Schuster, New York, 2006.

[38] Allen Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1992.

[39] Raymond Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems.* The MIT Press, Cambridge, Massachusetts, 2001.

[40] Raquel Ros, Josep Llus Arcos, Ramon Lopez de Mantaras, and Manuela M. Veloso. A case-based approach for coordinated action selection in robot soccer. *Artificial Intelligence*, 173(9–10):1014–1039, 2009.

[41] Alf Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941. Reprinted with minor changes, in *Philosophy of Science*, Vol. 11 (1944), pp. 30–46.

[42] Stuart J. Russell and Eric Wefald. *Do the Right Thing.* The MIT Press, Cambridge, Massachusetts, 1991.

[43] Scott Sanner and Craig Boutilier. Practical solution techniques for first-order MDPs. *Artificial Intelligence*, 173(5–6):748–788, 2009.

[44] Leonard Savage. *The Foundations of Statistics.* Dover, New York, 2 edition, 1972.

[45] Murray Shanahan. A logical formalisation of Ernie Davis' egg cracking problem. Unpublished manuscript, Imperial College London, 1997.

[46] Murray Shanahan. *Solving the Frame Problem.* The MIT Press, Cambridge, Massachusetts, 1997.

[47] Herbert A. Simon. *Administrative Behavior.* The Macmillan Company, New Yori, 1947.

[48] Herbert A. Simon and Jonathan Schaeffer. The game of chess. In Robert J. Aumann and Sergiu Hart, editors, *Handbook of Game Theory with Economic Applications, Vol. 1*, pages 1–17. North-Holland, Amsterdam, 1992.

[49] Brian Skyrms. *The Dynamics of Rational Deliberation.* Harvard University Press, Cambridge, Massachusetts, 1990.

[50] Robert C. Solomon. *The Passions: The Myth and Nature of Human Emotion.* Anchor

Press, New York, 1976.

[51] Richmond H. Thomason. Desires and defaults: A framework for planning with inferred goals. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 702–713, San Francisco, 2000. Morgan Kaufmann.

[52] Richmond H. Thomason. Preferences as conclusions. In Ulrich Junker, editor, *Preferences in AI and CP: Symbolic Approaches*, pages 94–98. AAAI Press, Menlo Park, California, 2002.

[53] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science, New Series*, 211(4481):453–458, 1981.

[54] Ubbo Visser and Hans-Dieter Burkhard. RoboCup: 10 years of achievements and future challenges. *The AI Magazine*, 28(2):115–132, 2007.

[55] Georg Henrik von Wright. Practical inference. *The Philosophical Review*, 72:159–179, 1963.

[56] Georg Henrik von Wright. *Practical Reason: Philosophical Papers, Volume 1.* Cornell University Press, Ithaca, 1983.

[57] Paul Weirich. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances.* Oxford University Press, Oxford, 2004.

[58] B.A.O. Williams. Imperative inference. *Analysis*, 23, Supplementary 1(3):30–36, 1963.

[59] Michael J. Wooldridge. *Reasoning about Rational Agents.* Cambridge University Press, Cambridge, England, 2000.

Received May 2014

A Note on Directions for Cumulativity

Philippe Besnard
CNRS, IRIT, Université de Toulouse, 31062 Toulouse Cedex, France
besnard@irit.fr

Abstract

Logical systems are often characterized as closure systems, by means of unary operators satisfying Reflexivity, Idempotence and Monotony. In order to capture non-monotone systems, Monotony can be replaced by Cumulativity, namely Restricted Cut and Cautious Monotony. This short note shows that in such a context, Restricted Cut is redundant.

Keywords: Non-Monotonic Consequence, Cumulativity.

1 Introduction

Tarski [7] introduced abtract logics, as consequence operations, later popularized by Scott [6] in connection with consequence relations à la Gentzen [3] (see also Gabbay [2]). The outcome is that abstract logics are often identified with closure operators over sets of formulas of a logical language. In symbols, given a logical language consisting of the set of formulas $\mathcal{F}$, a *consequence operator* is any C defined over $2^{\mathcal{F}}$ that satisfy all three axioms below:

$$\begin{array}{ll} X \subseteq C(X) & (Reflexivity) \\ C(C(X)) = C(X) & (Idempotence) \\ X \subseteq Y \Rightarrow C(X) \subseteq C(Y) & (Monotony) \end{array}$$

In an insightful attempt to also capture logical systems failing it, the last axiom was weakened by Makinson [5] into

$$X \subseteq Y \subseteq C(X) \Rightarrow C(X) = C(Y) \qquad (Cumulativity)$$

It was soon split into two “halves” dubbed Cautious Monotony and Restricted Cut.

$$\begin{array}{ll} X \subseteq Y \subseteq C(X) \Rightarrow C(X) \subseteq C(Y) & (\textit{Cautious Monotony}) \\ X \subseteq Y \subseteq C(X) \Rightarrow C(Y) \subseteq C(X) & (\textit{Restricted Cut}) \end{array}$$

It is the purpose of this note to show that, in its more natural context (Reflexivity and Idempotence), Cumulativity is not the combination of the two “halves” but is actually equivalent to one of them, namely Cautious Monotony.

Actually, it is shown below that if all three of Reflexivity, Idempotence, and Cautious Monotony hold, then so does Restricted Cut. In semantical terms, it means that $f(A) \subseteq A$ together with $f(f(A)) = f(A)$ and $f(A) \subseteq B \subseteq A \Rightarrow f(B) \subseteq f(A)$ give $f(A) \subseteq B \subseteq A \Rightarrow f(A) \subseteq f(B)$. Indeed, from $f(B) \subseteq f(A) \subseteq B \subseteq A$, it follows that $f(B) \subseteq f(A) \subseteq B$, so $f(A) = f(f(A)) \subseteq f(B)$.

2 The One Direction for Cumulativity

We switch to a sequent presentation (freely drawn upon Kleene [4] and Dummett [1]) as the result is less obvious there, although no less striking. That is, we consider a generalization of sequents

$$X_1, \ldots, X_n \vdash Y$$

such that the antecedent $X_1, \ldots, X_n$ consists of countably many formulas given as a (finite) series of sets for the sake of brevity. Thus, the rules have the following general form where W_i, X_j, Y, Z denote countable sets of formulas and *proviso* is a condition:

$$\frac{X_1, \ldots, X_n \vdash Y}{W_1, \ldots, W_m \vdash Z} \quad \textit{proviso}$$

Such a rule means that if *proviso* is true then $W_1, \ldots, W_m \vdash Z$ can be derived from $X_1, \ldots, X_n \vdash Y$.

The axioms, written here as rules with no premises, are meant to encode Reflexivity and have the following form

$$\frac{}{X \vdash Y} \quad (X \cap Y \neq \emptyset)$$

As for the rules, first please observe that we cannot include Left Thinning due to the motivation for Cumulativity. For the sake of brevity, we do not provide all rules and we resort to a rule mixing Left Interchange and Left Contraction as follows

Merging:

$$\frac{W, X, Y \vdash Z}{W, Y \vdash Z} \quad (X \subseteq Y)$$

Idempotence:

$$\frac{\{x \mid X \vdash x\} \vdash Z}{X \vdash Z}$$

In a monotone compact logic, the effect of Idempotence is in fact obtained through Restricted Cut. For such a logic, Idempotence is indeed an admissible rule in a sequent system with Restricted Cut (even as an admissible rule). However, Cumulativity (i.e., Restricted Cut and Cautious Monotony) was originally motivated by non-monotone logics hence postulating Idempotence makes sense when considering these logics.

Cautious Monotony:

$$\frac{X \vdash Z}{X, Y \vdash Z} \quad (y \in Y \Rightarrow X \vdash y)$$

Restricted Cut:

$$\frac{X, Y \vdash Z}{X \vdash Z} \quad (y \in Y \Rightarrow X \vdash y)$$

This formulation is intended to exhibit the fact that, under the same proviso, Cautious Monotony and Restricted Cut trigger converse inferences.

Theorem: *Restricted Cut is an admissible rule in any system enjoying Idempotence, Merging and Cautious Monotony.*

Proof. 1. Applying Cautious Monotony ($X = \Gamma$ and $Y = \Theta$)

$$\frac{\Gamma \vdash \Delta}{\Gamma, \Theta \vdash \Delta} \quad (\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$$

Stated otherwise, if $(\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$ then $(\gamma \in \{\gamma \mid \Gamma \vdash \gamma\} \Rightarrow \Gamma, \Theta \vdash \gamma)$.

2. Applying Cautious Monotony ($X = \Gamma, \Theta$ and $Y = \{\gamma \mid \Gamma \vdash \gamma\}$)

$$\frac{\Gamma, \Theta \vdash \Delta}{\Gamma, \Theta, \{\gamma \mid \Gamma \vdash \gamma\} \vdash \Delta} \quad (\gamma \in \{\gamma \mid \Gamma \vdash \gamma\} \Rightarrow \Gamma, \Theta \vdash \gamma)$$

3. In view of what we just proved in Step 1, we obtain

$$\frac{\Gamma, \Theta \vdash \Delta}{\Gamma, \Theta, \{\gamma \mid \Gamma \vdash \gamma\} \vdash \Delta} \quad (\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$$

4. Trivially, if $(\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$ then $\Theta \subseteq \{\gamma \mid \Gamma \vdash \gamma\}$ hence applying Merging gives

$$\frac{\Gamma, \Theta \vdash \Delta}{\Gamma, \{\gamma \mid \Gamma \vdash \gamma\} \vdash \Delta} \quad (\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$$

5. According to the axioms, $\Gamma \subseteq \{\gamma \mid \Gamma \vdash \gamma\}$ hence applying Merging again gives

$$\frac{\Gamma, \Theta \vdash \Delta}{\{\gamma \mid \Gamma \vdash \gamma\} \vdash \Delta} \quad (\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$$

6. Lastly, applying Idempotence yields

$$\frac{\Gamma, \Theta \vdash \Delta}{\Gamma \vdash \Delta} \quad (\theta \in \Theta \Rightarrow \Gamma \vdash \theta)$$

   which is exactly Restricted Cut.

□

The author is aware of not being the first to figure all this out but after discussing with various colleagues working in the field, it appears that this was largely ignored, and unpublished to the best of his knowledge, hence it could justify a brief note such as the present one.

3 Conclusion

Reflexivity and Idempotence are the most desirable features of a logical system. If Monotony is to be weakened, Cumulativity conveys the attractive idea that intermediate conclusions could be, as premises, freely added or freely removed without changing the overall set of conclusions. However, we have shown that Cumulativity is, unexpectedly, captured by part of the idea, Cautious Monotony, although the latter only imposes that intermediate conclusions could be removed from the premises with no loss among conclusions. In other words, the other part of the idea, Restricted Cut, is actually otiose with respect to weakening Monotony, or Full Cut for that matter. Is there a context in which Restricted Cut would play some logical role? (Please observe that a formal role is possible, together with Reflexivity, as Idempotence ensues.)

References

[1] M. Dummett. *Elements of Intuitionism.* Oxford University Press, 2nd edition, 2000.

[2] D. Gabbay. *Semantical Investigations in Heyting's Intuitionistic Logic*, chapter Consequence Relations. Reidel, Dordrecht, 1981.

[3] G. Gentzen. *The Collected Works of Gerhard Gentzen (M. E. Szabo, editor)*, chapter Investigations into Logical Deduction. North-Holland, Amsterdam, 1969.

[4] S. Kleene. *Introduction to Metamathematics.* North-Holland, Amsterdam, 1952.

[5] D. Makinson. *Handbook of Logic in Artificial Intelligence and Logic Programming (D. M. Gabbay, C. J. Hogger, and J. A. Robinson (editors)*, volume III (D. Nute, volume co-ordinator), chapter General Patterns in Nonmonotonic Reasoning, Clarendon Press, Oxford, 1994.

[6] D. Scott. Completeness and Axiomatizability in Many-Valued Logic, *Proceedings of the Tarski Symposium (L. Henkin, editor)*, AMS, Providence, pp. 411-435, 1974.

[7] A. Tarski. *Logic, Semantics, Metamathematics (E. H. Woodger, editor)*, chapter On Some Fundamental Concepts of Metamathematics, Oxford University Press, 1956.

Received March 2014

A Simple and Complete Model Theory for Intensional and Extensional Untyped λ-Equality

Michael Gabbay
Department of Philosophy, Kings College London

Murdoch Gabbay
Department of Computer Science, Heriott-Watt University, Edinburgh

Abstract

We present a sound and complete model theory for theories of β-reduction with or without η-expansion. The models of this paper derive from structures of modal logic: we use ternary accessibility relations on 'possible worlds' to model the action of intensional and extensional lambda-abstraction in much the same way binary accessibility relations are used to model the box operators of a normal multi-modal logic.

Keywords: Lambda Calculus, Reduction, Intensional Equality, Extensional Equality, Model Theory, Completeness, Kripke Frames, Possible World Semantics, Modal Logic.

1 Introduction

We extend the method of [6] by which we interpret λ-terms compositionally on 'possible world' structures. The simplicity of the structures is striking, moreover, they provide a surprising richness of interpretations of function abstraction and application.

Our primary goal is to show how the models can differentiate between *extensional* and *intensional* λ-equality, and provide semantic characterisation (i.e. completeness) theorems for both. We shall then hint at how richer λ-languages can be interpreted.

We are very grateful to the input and suggestions of the referees of this journal paper and of its previous incarnation as a conference paper at CiE 2012.

The key idea in this paper is a class of models, presented in Section 2.2, although an important syntactic consideration is required first in Section 2.1. These ideas bear some similarity to the reduction models of [18] in that they get us as far as λ-reduction only. Then, in Section 4 we use the results of the earlier sections to provide a characterisation theorem for λ-equality (both with and without η-equality, i.e. extensional and intensional).

2 The models, computation, logic

2.1 The language and logic

Definition 2.1. Fix a countably infinite set of **variables**.

Define a language $\mathcal{L}_\lambda$ of **λ-terms** by: $\mathtt{t} \ ::= \ \mathtt{x} \mid \lambda\mathtt{x}.\mathtt{t} \mid \mathtt{t}{\cdot}\mathtt{t}$
$\lambda\mathtt{x}$ binds in $\lambda\mathtt{x}.\mathtt{t}$. For example, $\mathtt{x}$ is bound (not free) in $\lambda\mathtt{x}.\mathtt{x}{\cdot}\mathtt{y}$.

We write $\mathtt{t}[\mathtt{x}/\mathtt{s}]$ for the usual capture-avoiding substitution. For example, $(\lambda\mathtt{z}.\mathtt{y})[\mathtt{y}/\mathtt{x}] = \lambda\mathtt{z}.\mathtt{x}$, and $(\lambda\mathtt{x}.\mathtt{y})[\mathtt{y}/\mathtt{x}] = \lambda\mathtt{z}.\mathtt{x}$ where $\mathtt{z}$ is an arbitrary fresh variable. If $\mathtt{x}_1 \dots \mathtt{x}_n$ is a sequence of variables and $\mathtt{t}_1 \dots \mathtt{t}_n$ is an equally long sequence of terms then we write $\mathtt{t}[\mathtt{x}_i/\mathtt{t}_i]$ for the simultaneous substitution in $\mathtt{t}$ of each $\mathtt{x}_i$ by its corresponding $\mathtt{t}_i$.

We write $\mathtt{t}[\mathtt{x}{:}{-}\mathtt{s}]$ for the (unusual) non-capture avoiding substitution. For example, $(\lambda\mathtt{x}.\mathtt{x})[\mathtt{x}{:}{-}\mathtt{y}] = \lambda\mathtt{x}.\mathtt{y}$, and $(\lambda\mathtt{x}.\mathtt{y})[\mathtt{y}{:}{-}\mathtt{x}] = \lambda\mathtt{x}.\mathtt{x}$

We now turn to λ-reduction. It is important for us to consider not merely the relation of λ-reduction, but a relation of λ-reduction *with assumptions.* We therefore need to define some basic, and familiar, rules of λ-reduction but allow for a set of assumed additional reductions.

Remark 2.2. We shall define a basic relation on terms that follows the familiar reduction rule of β-contraction (Definition 2.3). To help with the completeness theorem of Section 3 we will need to consider a conservative extension of the familiar λ-calculus (Definition 2.5). To facilitate the proof that this extension really is conservative (Theorem 2.8), we present the λ-calculus in the non-axiomatic style of [12, Def. 1.24].

Definition 2.3. Let Γ be a set of pairs of terms of $\mathcal{L}_\lambda$. We define a **reduction relation** $\longrightarrow_\Gamma$ on terms of $\mathcal{L}_\lambda$ using Figure 1. A **derivation** is a sequence of terms $\mathtt{t}_1, \dots, \mathtt{t}_n$ such that $\mathtt{t}_i \longrightarrow_\Gamma \mathtt{t}_{i+1}$ for each $1 \leq i < n$.

Remark 2.4. If $\Gamma = \varnothing$ then $\longrightarrow_\Gamma$ is the familiar relation of untyped β-reduction.

Let $\mathtt{x}$ occur free *only once* in $\mathtt{t}$. Let $\mathtt{x}_1 \dots \mathtt{x}_n$ be any sequence of variables and $\mathtt{t}_1 \dots \mathtt{t}_n$ be any (equally long) sequence of terms.

$$\begin{array}{lrcll}
(\beta) & \mathtt{t}[\mathtt{x}{:-}(\lambda\mathtt{x}.\mathtt{s})\cdot\mathtt{r}] & \longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{x}/\mathtt{r}]] & \\
(\mathit{ass}) & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{x}_i/\mathtt{t}_i]] & \longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{r}[\mathtt{x}_i/\mathtt{t}_i]] & (\langle\mathtt{s},\mathtt{r}\rangle \in \Gamma) \\
(\alpha) & \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.\mathtt{s}] & \longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{z}.\mathtt{s}[\mathtt{y}/\mathtt{z}]] &
\end{array}$$

The rule (ass) says that any (capture avoiding) substitution instance of $\mathtt{s}$ may be replaced in any $\mathtt{t}$, without worrying about variable capture, by its matching substitution instance of $\mathtt{r}$.[1]

Figure 1: λ-reduction for $\mathcal{L}_\lambda$

Let $\mathtt{x}$ occur free *only once* in $\mathtt{t}$. Let $\mathtt{x}_1 \dots \mathtt{x}_n$ be any sequence of variables and $\mathtt{t}_1 \dots \mathtt{t}_n$ be any (equally long) sequence of terms.

$$\begin{array}{lrcll}
(\beta) & \mathtt{t}[\mathtt{x}{:-}(\lambda\mathtt{x}.\mathtt{s})\cdot\mathtt{r}] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{x}/\mathtt{r}]] & \\
(\mathit{ass}) & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{x}_i/\mathtt{t}_i]] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{r}[\mathtt{x}_i/\mathtt{t}_i]] & \langle\mathtt{s},\mathtt{r}\rangle \in \Gamma \\
(\alpha) & \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.\mathtt{s}] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{z}.\mathtt{s}[\mathtt{y}/\mathtt{z}]] & \\
(\beta^*) & \mathtt{t}[\mathtt{x}{:-}(\lambda\mathtt{x}.\mathtt{s}) * \mathtt{r}] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{x}/\mathtt{r}]] & \\
(\mathit{sub}) & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}\cdot\mathtt{r}] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\mathtt{s} * \mathtt{r}] & \\
(\eta^*) & \mathtt{t}[\mathtt{x}{:-}\mathtt{s}] & \Longrightarrow_\Gamma & \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})] & (\mathtt{y} \text{ not free in } \mathtt{t})
\end{array}$$

Figure 2: λ-reduction for $\mathcal{L}^*_\lambda$

Definition 2.5. Define $\mathcal{L}^*_\lambda$ by: $\quad \mathtt{t} \ ::= \ \mathtt{x} \mid \lambda\mathtt{x}.\mathtt{t} \mid \mathtt{t}\cdot\mathtt{t} \mid \mathtt{t} * \mathtt{t}$

Definition 2.6. Let Γ be a set of pairs of terms of $\mathcal{L}_\lambda$ (*not* $\mathcal{L}^*_\lambda$). Define a reduction relation $\Longrightarrow_\Gamma$ on terms of $\mathcal{L}^*_\lambda$ using Figure 2. Again, a **derivation** is a sequence of terms $\mathtt{t}_1, \dots, \mathtt{t}_n$ such that $\mathtt{t}_i \Longrightarrow_\Gamma \mathtt{t}_{i+1}$ for each $1 \leq i < n$.

Remark 2.7. Notice that we do not allow terms unique to $\mathcal{L}^*_\lambda$ to be assumptions in derivations. This is because the paper is concerned with characterising reduction and equality only in the more familiar language $\mathcal{L}_\lambda$, and $\mathcal{L}^*_\lambda$ is merely a means to that end. Allowing assumed reductions for $\mathcal{L}^*_\lambda$ causes problems for the Theorem 2.8.

Theorem 2.8. *If* $\mathtt{t}_1$ *and* $\mathtt{t}_2$ *are terms of* $\mathcal{L}_\lambda$ *then* $\mathtt{t}_1 \Longrightarrow_\Gamma \mathtt{t}_2$ *implies* $\mathtt{t}_1 \longrightarrow_\Gamma \mathtt{t}_1$. *In other words* $\Longrightarrow_\Gamma$ *is* **conservative** *over* $\longrightarrow_\Gamma$.

[1] So for example if $\langle\mathtt{x}, \lambda\mathtt{y}.(\mathtt{x}\cdot\mathtt{y})\rangle \in \Gamma$ then $\longrightarrow_\Gamma$ is a reduction relation allowing η-expansion: $\mathtt{s}$ may be rewritten, inside any term $\mathtt{t}$, to $\lambda\mathtt{y}.(\mathtt{s}\cdot\mathtt{y})$ provided the lambda-operator $\lambda\mathtt{y}$ does not bind in $\mathtt{s}$ (although variables in $\mathtt{s}$ may be bound by abstractions in the wider context $\mathtt{t}$.

Proof. Suppose that $\mathtt{t}_1 \Longrightarrow_\Gamma \mathtt{t}_2$, where $\mathtt{t}_1, \mathtt{t}_2 \in \mathcal{L}_\lambda$. We argue that any such derivation can be converted into a derivation that $\mathtt{t}_1 \longrightarrow_\Gamma \mathtt{t}_2$.

We first argue that any application of (η) or (sub) can be pushed after an application of any other rule or eliminated entirely.

- Suppose we have the following derivation segment:

$$\mathtt{t}[\mathtt{x}{:-}\mathtt{s}] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})] \Longrightarrow_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s}' * \mathtt{y})]$$

where $\mathtt{s}'$ is derived from $\mathtt{s}$ by an application of any rule, then we may easily swap the rule applications:

$$\mathtt{t}[\mathtt{x}{:-}\mathtt{s}] \Longrightarrow_\Gamma \mathtt{t}[\mathtt{x}{:-}\mathtt{s}'] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s}' * \mathtt{y})]$$

- The cases where $\mathtt{t}'[\mathtt{x}{:-}\mathtt{s}]$ is derived from $\mathtt{t}[\mathtt{x}{:-}\mathtt{s}]$ is similar. For example:

$$\mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}[\mathtt{x}{:-}\mathtt{s}]\big] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})]\big] \overset{(ass)}{\Longrightarrow}_\Gamma \mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}'[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})]\big]$$

may be replaced, given that Γ contains only terms from $\mathcal{L}_\lambda$,[2] by:

$$\mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}[\mathtt{x}{:-}\mathtt{s}]\big] \overset{(ass)}{\Longrightarrow}_\Gamma \mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}'[\mathtt{x}{:-}\mathtt{s}]\big] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}\big[\mathtt{z}{:-}\mathtt{r}'[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})]\big]$$

There are also the following three special cases:

$$\begin{gathered}\mathtt{t}[\mathtt{x}{:-}\mathtt{s}{\cdot}\mathtt{r}] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y}){\cdot}\mathtt{r}] \overset{(\beta)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\mathtt{s}{\cdot}\mathtt{r}]\\ \text{becomes}\\ \mathtt{t}[\mathtt{x}{:-}\mathtt{s}{\cdot}\mathtt{r}]\end{gathered}$$

$$\begin{gathered}\mathtt{t}[\mathtt{x}{:-}\mathtt{s} * \mathtt{r}] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y}) * \mathtt{r}] \overset{(\beta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\mathtt{s} * \mathtt{r}]\\ \text{becomes}\\ \mathtt{t}[\mathtt{x}{:-}\mathtt{s} * \mathtt{r}]\end{gathered}$$

$$\begin{gathered}\mathtt{t}[\mathtt{x}{:-}\mathtt{s}] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}.(\mathtt{s} * \mathtt{y})] \overset{(\alpha)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{z}.(\mathtt{s} * \mathtt{z})]\\ \text{becomes}\\ \mathtt{t}[\mathtt{x}{:-}\mathtt{s}] \overset{(\eta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{z}.(\mathtt{s} * \mathtt{z})]\end{gathered}$$

[2] And so the application of (ass) cannot depend on the preceding application of (η^*).

- By a similar reasoning it follows that (sub) can be pushed in front of any other rule, with the exception of the special case of the derivation segment

$$\mathtt{t}[\mathtt{x}{:-}\lambda \mathtt{y}.\mathtt{s}{\cdot}\mathtt{r}] \overset{(\mathit{sub})}{\Longrightarrow}_{\Gamma} \mathtt{t}[\mathtt{x}{:-}\lambda \mathtt{y}.\mathtt{s} * \mathtt{r}] \overset{(\beta^*)}{\Longrightarrow}_{\Gamma} \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{y}/\mathtt{r}]]$$

which may be replaced by:

$$\mathtt{t}[\mathtt{x}{:-}\lambda \mathtt{y}.\mathtt{s}{\cdot}\mathtt{r}] \overset{(\beta)}{\Longrightarrow}_{\Gamma} \mathtt{t}[\mathtt{x}{:-}\mathtt{s}[\mathtt{y}/\mathtt{r}]]$$

It follows that any derivation may be replaced by a derivation where the last application is an instance of (η) or (sub), if either appears in the derivation at all. Since (η) and (sub) introduce an instance of $*$, if $\mathtt{t}_2 \in \mathcal{L}_\lambda$ then no instances of (η) or (sub) occur in the derivation. Furthermore, since $\mathtt{t}_1 \in \mathcal{L}_\lambda$ it follows that the derivation contains no instances of (β^*) or occurrences of $*$. This implies that $\mathtt{t}_1 \longrightarrow_\Gamma \mathtt{t}_2$. □

2.2 Frames and interpreting λ-terms

Given Theorem 2.8 we will work with $\mathcal{L}^*_\lambda$.

Definition 2.9. If W is a set, write $\mathcal{P}(W)$ for the set of subsets of W.

An **intensional frame** F is a 4-tuple $(W, \bullet, \circ, H)$ where:

- W a set of **worlds**,
- $\bullet$ and $\circ$ are functions from $W \times W$ to $\mathcal{P}(W)$ such that $\bullet \subseteq \circ$.
- $H \subseteq \mathcal{P}(W)$.

Remark 2.10. Subsets of W will serve as denotations of λ-terms (Definition 2.13) and $H \subseteq \mathcal{P}(W)$ ('H' for 'Henkin') plays a similar role to the structure of Henkin models for higher-order logic [2, 11, 19]. This makes our completeness results possible and is a famous issue for second- and higher-order logics: powersets are too large and for completeness results to be possible we must cut them down — at least when we quantify. This is why in Definition 2.13, the binders restrict quantification from $\mathcal{P}(W)$ down to H.

The reader familiar with modal logic can think of $\bullet$ and $\circ$ as ternary 'accessibility relations' $R_\bullet$ and $R_\circ$ such that $R_\bullet w_1 w_2 w_3$ if and only if $w_3 \in w_1 \bullet w_2$ (and similarly for $R_\circ$). We can also think of $\bullet$ and $\circ$ as non-deterministic 'application' operations, but note that intensional frames are not applicative structures — an applicative structure would map $W \times W$ to W, whereas in the case of intensional frames, $W \times W$ maps to $\mathcal{P}(W)$.

Definition 2.11. Let $F = (W, \bullet, \circ, H)$ be an intensional frame and $S_1, S_2 \subseteq W$ and $w \in W$. Then the functions $\bullet$ and $\circ$ induce functions from $W \times \mathcal{P}(W)$ and $\mathcal{P}(W) \times \mathcal{P}(W)$ to $\mathcal{P}(W)$ by: $w \bullet S = \bigcup\{w \bullet w' \mid w' \in S\}$ and $S_1 \bullet S_2 = \bigcup\{w_1 \bullet w_2 \mid w_1 \in S_1,\ w_2 \in S_2\}$ (and similarly for $\circ$).

Definition 2.12. Suppose $F = (W, \bullet, \circ, H)$ is a frame. A **valuation** (to F) is a map from variables to sets of worlds (elements of $\mathcal{P}(W)$) that are in H. v will range over valuations.

If $\mathtt{x}$ is a variable, $h \in H$, and v is a valuation, then write $v[\mathtt{x} \mapsto h]$ for the valuation mapping $\mathtt{x}$ to h and mapping $\mathtt{y}$ to $v(\mathtt{y})$ for any other $\mathtt{y}$.

Definition 2.13. Define an **denotation** of $\mathtt{t}$ inductively by:

$$[\![\mathtt{x}]\!]^v = v(\mathtt{x}) \qquad [\![\mathtt{t{\cdot}s}]\!]^v = [\![\mathtt{t}]\!]^v \bullet [\![\mathtt{s}]\!]^v \qquad [\![\mathtt{t} * \mathtt{s}]\!]^v = [\![\mathtt{t}]\!]^v \circ [\![\mathtt{s}]\!]^v$$

$$[\![\lambda\mathtt{x}.\mathtt{t}]\!]^v = \{w \mid w \circ h \subseteq [\![\mathtt{t}]\!]^{v[\mathtt{x}\mapsto h]} \text{ for all } h \in H\}$$

Remark 2.14. By elementary set theory: $[\![\lambda\mathtt{x}.\mathtt{t}]\!]^v = \bigcap_{h\in H}\{w \mid w \circ h \subseteq [\![\mathtt{t}]\!]^{v[\mathtt{x}\mapsto h]}\}$

We are particularly interested in frames where the denotation of every λ-term is a member of H. This is because Definition 2.13 interprets λ as a kind of quantifier over all members of H. β-reduction is then valid analogously to universal instantiation in first order logic ($\forall x.Fx \vDash Ft$),[3] and so requires that every possible instantiation (i.e. every term denotation) is a member of H.

Remark 2.15. Consider the definition of application and abstraction in a graph model with carrier set $\mathcal{P}(A)$ (Scott semantics), where $\mapsto\colon \mathcal{P}_{fin}(A) \times A \mapsto A$ is an arbitrary injective map and $v : Var \mapsto \mathcal{P}(A)$ is an arbitrary environment:

$$\begin{aligned} X \bullet Y &= \{\alpha \mid (\exists a \subseteq_{fin} A) \text{ s.t. } a \mapsto \alpha \in X \text{ and } a \subseteq Y\} && (X, Y \subseteq A) \\ [\![\lambda\mathtt{x}.\mathtt{t}]\!]^v &= \{a \mapsto \alpha \mid \alpha \in [\![\mathtt{t}]\!]^{v[x\mapsto a]}\} && (\mathtt{t} \text{ a lambda term}) \\ &= \{a \mapsto \alpha \mid (\forall h \in \mathcal{P}(A))\{a \mapsto \alpha\} \bullet h \subseteq [\![\mathtt{t}]\!]^{v[x\mapsto h]}\} \end{aligned}$$

Very roughly speaking (our construction is more general), the above definitions of application and abstraction in graph models are abstracted in this paper as follows, where H is a fixed subset of $\mathcal{P}(A)$:

$$\begin{aligned} \bullet : H \times H \mapsto H &= \text{any linear map in both arguments} \\ [\![\lambda\mathtt{x}.\mathtt{t}]\!]^v &= \{\alpha \mid (\forall h \in H)\{\alpha\} \bullet h \subseteq [\![\mathtt{t}]\!]^{v[x\mapsto h]}\} \end{aligned}$$

where $v : Var \mapsto H$ is an H-environment. This semantics is of perhaps of additional interest because it does not codify the step functions of continuous semantics.

[3] Perhaps a better analogy would be $\forall x.(Fx \to Gx) \wedge Ft \vDash Gt$, where conjunctions corresponds to $\cdot$ and the quantified expression corresponds to a λ-term.

Lemma 2.16. *1. If* $\mathtt{x}$ *is not free in* $\mathtt{t}$*, then for any* $h \in H$*,* $[\![\mathtt{t}]\!]^v = [\![\mathtt{t}]\!]^{v[\mathtt{x}\mapsto h]}$*.*
2. $[\![\mathtt{t}[\mathtt{x}/\mathtt{s}]]\!]^v = [\![\mathtt{t}]\!]^{v[\mathtt{x}\mapsto[\![\mathtt{s}]\!]^v]}$

Proof. Both parts follow by easy inductions on $\mathtt{t}$. □

Definition 2.17. A frame is **faithful** when for every v and every $\mathtt{t} \in \mathcal{L}_\lambda$, $[\![\mathtt{t}]\!]^v \in H$. That is, a frame is faithful when H contains the interpretation of every λ term in $\mathcal{L}_\lambda$ independently of v.

Remark 2.18. Definition 2.17 is not ideal. A semantic characterisation of faithfulness — as a condition on H, $\bullet$ and $\circ$ — is desirable. We cannot present such a charactersiation in this paper except in the special cases of theories of β-equality which we defer until 5.1 of Section 5. In spite of this, the structures characterised by 2.17 are informative because they allow us to break down λ-abstraction into a quantification over worlds with a ternary accessibility relation; the denotations of λ-terms then become simply sets of worlds. On this analysis the domain of quantification for λ-abstraction — the actual set of denotations of λ-terms — is H, a subset of the set of all possible denotations $\mathcal{P}(W)$. This is a common pattern, for example in topological semantics for modal logics and for intuitionistic logic the set of denotations, the analogue of H, is the set of all open subsets of the domain rather than the powerset itself. It is then no accident that H will reveal further significance: it will be useful later in characterising the difference between extensional and intensional λ-calculus (see Remark 3.18). Unlike in the modal case, we offer no general way in this paper of characterising H beyond simply saying that it must contain $[\![\mathtt{t}]\!]^v$ for each valuation v and term $\mathtt{t}$. For a considerably more complex characterisation of H in topological terms see [8].

Lemma 2.19. *If we interpret* $\Longrightarrow_\Gamma$ *as subset inclusion then all the rules of Figure 2 are sound for faithful intensional frames.*

Proof. By routine calculations from the definitions. We show only (β) and (η^*) here, the others are equally straightforward.

$$
\begin{array}{lll}
\llbracket \lambda \mathtt{x}.\mathtt{t}{\cdot}\mathtt{s} \rrbracket^v & & \\
 & = \llbracket \lambda \mathtt{x}.\mathtt{t} \rrbracket^v \bullet \llbracket \mathtt{s} \rrbracket^v & \text{Definition 2.13} \\
 & = \bigcap_{h \in H} \{ w \mid w \circ h \subseteq \llbracket \mathtt{t} \rrbracket^{v[\mathtt{x} \mapsto h]} \} \bullet \llbracket \mathtt{s} \rrbracket^v & \text{Definition 2.13} \\
 & \subseteq \{ w \mid w \circ \llbracket \mathtt{s} \rrbracket^v \subseteq \llbracket \mathtt{t} \rrbracket^{v[\mathtt{x} \mapsto \llbracket \mathtt{s} \rrbracket^v]} \} \bullet \llbracket \mathtt{t} \rrbracket^v & \llbracket \mathtt{s} \rrbracket^v \in H \\
 & \subseteq \{ w \mid w \bullet \llbracket \mathtt{s} \rrbracket^v \subseteq \llbracket \mathtt{t} \rrbracket^{v[\mathtt{x} \mapsto \llbracket \mathtt{s} \rrbracket^v]} \} \bullet \llbracket \mathtt{t} \rrbracket^v & \bullet \subseteq \circ \\
 & \subseteq \llbracket \mathtt{t} \rrbracket^{v[\mathtt{x} \mapsto \llbracket \mathtt{s} \rrbracket^v]} & \text{Definition 2.11} \\
 & = \llbracket \mathtt{t}[\mathtt{x}/\mathtt{s}] \rrbracket^v & \text{Lemma 2.16}
\end{array}
$$

$$
\begin{array}{lll}
\llbracket \mathtt{t} \rrbracket^v & \subseteq \bigcap_{h \in H} \{ w \mid w \circ h \subseteq \llbracket \mathtt{t} \rrbracket^v \circ h \} & \text{Definition 2.11} \\
 & = \bigcap_{h \in H} \{ w \mid w \circ h \subseteq \llbracket \mathtt{t} * \mathtt{x} \rrbracket^{v[\mathtt{x} \mapsto h]} \} & \mathtt{x}\text{ not free in } \llbracket \mathtt{t} \rrbracket^v \\
 & = \llbracket \lambda \mathtt{x}.(\mathtt{t} * \mathtt{x}) \rrbracket^v & \text{Definition 2.13}
\end{array}
$$

□

2.3 Soundness

Definition 2.20. – A **model** M is a pair $\langle F, v \rangle$ where F is a faithful intensional frame and v is a valuation on F such that $v(\mathtt{t}) \in H \in F$ for every $\mathtt{t}$.

– A frame F is **Γ-sensitive** if $\llbracket \mathtt{t} \rrbracket^v \subseteq \llbracket \mathtt{s} \rrbracket^v$ for every v and every $\langle \mathtt{t}, \mathtt{s} \rangle \in \Gamma$.

– A model $\langle F, v \rangle$ is **Γ-sensitive** if F is Γ-sensitive.

Remark 2.21. We could have defined a model as a pair $\langle F, v \rangle$ where F is a (possibly not faithful) frame and v is a valuation on F such that $\llbracket \mathtt{t} \rrbracket^v \in H \in F$ for every $\mathtt{t}$. But since the completeness theorem 3.12 holds for the stronger notion of a model we shall use that. Intuitively a Γ-sensitive frame or model can be thought of as giving $\langle \mathtt{t}, \mathtt{s} \rangle \in \Gamma$ the meaning that however the variables of $\mathtt{t}$ and $\mathtt{s}$ are interpreted, $\mathtt{t}$'s denotation is a subset of $\mathtt{s}$'s.

Remark 2.22. This paper approaches lambda calculus from the angle of modal logic and so we retain the 'normal' practice of describing the model theory in terms of frames and models: a frame is sufficient to fix the interpretation of the closed terms, and a model interprets the open terms (e.g. as in [10]. This also matches the 'normal' practice in the model theory of first order logic of distinguishing a structure from a model – a structure together with a variable assignment – as in [3]. This differs from the 'normal' terminology for lambda calculus which would use the term 'model' to refer to our frames (e.g. in [1]).

Lemma 2.23. $\bullet$ *and* $\circ$ *are* **monotone**. *That is, $h_1 \subseteq h_2$ implies $h \bullet h_1 \subseteq h \bullet h_2$ and $h_1 \bullet h \subseteq h_2 \bullet h$ for any h, and similarly for $\circ$.*

Proof. By the pointwise definitions of $\bullet$ and $\circ$. For example:

$$\begin{array}{rcll} h \bullet h_1 & = & \bigcup\{w \bullet w_1 \mid w \in h \text{ and } w_1 \in h_1\} & \text{Def. 2.11} \\ & \subseteq & \bigcup\{w \bullet w_1 \mid w \in h \text{ and } w_1 \in h_2\} & \text{if } h_1 \subseteq h_2 \end{array}$$

□

Lemma 2.24. *If* $[[\mathtt{s}]]^v \subseteq [[\mathtt{r}]]^v$ *for all* v *on some faithful* F*, then for any* v *on* F $[[\mathtt{t}[\mathtt{x}{:-}\mathtt{s}]]]^v \subseteq [[\mathtt{t}[\mathtt{x}{:-}\mathtt{r}]]]^v$

Proof. By induction on $\mathtt{t}$.

– If $\mathtt{t}$ is a variable the result is easy.

– If $\mathtt{t}$ is $\mathtt{t}_1{\cdot}\mathtt{t}_2$ or $\mathtt{t}_1 * \mathtt{t}_2$ then the result follows from the induction hypothesis and the monotonicity of $\bullet$ and $\circ$ (Lemma 2.23).

– If $\mathtt{t}$ is $\lambda\mathtt{x}\mathtt{t}'$, then $\mathtt{t}[\mathtt{x}{:-}\mathtt{s}]$ is $\lambda\mathtt{x}.\mathtt{t}'[\mathtt{x}{:-}\mathtt{s}]$. And so:

$$\begin{array}{rcll} [[\lambda\mathtt{x}.\mathtt{t}'[\mathtt{x}{:-}\mathtt{s}]]]^v & = & \bigcap_{h\in H}\{w \mid w \circ h \subseteq [[\mathtt{t}'[\mathtt{x}{:-}\mathtt{s}]]]^{v[\mathtt{x}\mapsto h]}\} & \text{Def. 2.13} \\ & \subseteq & \bigcap_{h\in H}\{w \mid w \circ h \subseteq [[\mathtt{t}'[\mathtt{x}{:-}\mathtt{r}]]]^{v[\mathtt{x}\mapsto h]}\} & \text{Ind. Hyp} \\ & = & [[\lambda\mathtt{x}.\mathtt{t}'[\mathtt{x}{:-}\mathtt{r}]]]^v & \text{Def. 2.13} \end{array}$$

And the argument is similar if $\mathtt{t}$ is $\lambda\mathtt{y}.\mathtt{t}'$ for $\mathtt{y} \neq \mathtt{x}$

□

Theorem 2.25. $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ *implies* $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ *in all* Γ*-sensitive (faithful) models* M*.*

Proof. Theorem 2.19 entails that each rule of Figure 2 holds in all models, and by definition, if $\langle\mathtt{t},\mathtt{s}\rangle \in \Gamma$ then $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ for all v in any Γ-sensitive model. The result then follows by Lemma 2.24. □

3 Completeness for λ-reduction

Ultimately, we wish to show that if $\mathtt{t} \not\twoheadrightarrow_\Gamma \mathtt{s}$ then there is a Γ-sensitive model M (Def. 2.20) where $[[\mathtt{t}]]^v \nsubseteq [[\mathtt{s}]]^v$. We first show that $\mathtt{t} \not\Longrightarrow_\Gamma \mathtt{s}$ implies such an M exists if $\mathtt{t}, \mathtt{s} \in \mathcal{L}_\lambda$, and then we appeal to Theorem 2.8.

First we form the languages $\mathcal{L}_{\lambda_c}, \mathcal{L}^*_{\lambda_c}$ by adding infinitely many new constant symbols $\mathtt{c}_1, \mathtt{c}_2 \ldots$ to $\mathcal{L}_\lambda$ and $\mathcal{L}^*_{\lambda_c}$. Since the language is countable we can enumerate its terms $\mathtt{t}_1, \mathtt{t}_2 \ldots$, which may contain the new constants, and the new constants alone $\mathtt{c}_1, \mathtt{c}_2 \ldots$. We describe a one-one function f from terms to constants.

> $f(\mathtt{t}_i) = \mathtt{c}_j$ where j is the least number such that $j > i$ and $\mathtt{c}_j$ does not occur in $\mathtt{t}_i$ nor is the value under f of any $\mathtt{t}_k$ for $k < i$.

Thus f is a one-one function that assigns a distinct 'fresh' constant to each term of the language, so $f(\mathtt{t})$ is a constant that 'names' $\mathtt{t}$. These play the role of witness constants in the construction of the canonical frame in Theorem 3.8. The $f(\mathtt{t})$ also help us carry out inductions on the size of λ-terms, as $\mathtt{t}[\mathtt{x}/f(\mathtt{s})]$ is smaller than $\lambda\mathtt{x}.\mathtt{t}$ even if $\mathtt{t}[\mathtt{x}/\mathtt{s}]$ might not be.

Definition 3.1. Define a reduction relation $\Longrightarrow^f_\Gamma$ on terms of of $\mathcal{L}^*_{\lambda_c}$ by setting $\mathtt{t} \Longrightarrow^f_\Gamma \mathtt{s}$ if $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ and using the rule:

$$(con) \quad \begin{array}{r} \mathtt{t}[\mathtt{x}/\mathtt{s}] \Longrightarrow^f_\Gamma \mathtt{t}[\mathtt{x}/f(\mathtt{s})] \\ \mathtt{t}[\mathtt{x}/f(\mathtt{s})] \Longrightarrow^f_\Gamma \mathtt{t}[\mathtt{x}/\mathtt{s}] \end{array}$$

In other words, $\Longrightarrow^f_\Gamma$ extends $\Longrightarrow_\Gamma$ with the rule (con), which makes $\mathtt{t}$ and its corresponding $f(\mathtt{t})$ inter-reducible.

Remark 3.2. Simply extending $\Longrightarrow_\Gamma$ by insisting that $\{\langle \mathtt{t}, f(\mathtt{t})\rangle, \langle f(\mathtt{t}), \mathtt{t}\rangle\} \subseteq \Gamma$ for every $\mathtt{t}$ is not equivalent to defining $\Longrightarrow^f_\Gamma$ as we have done above. For example, consider the individual variable $\mathtt{x}$: if $f(\mathtt{x})$ is $\mathtt{c}$ and $\langle \mathtt{x}, \mathtt{c}\rangle \in \Gamma$ then by (ass), $\mathtt{t} \Longrightarrow_\Gamma \mathtt{c}$ for any $\mathtt{t}$.

Lemma 3.3. *If* $\mathtt{t} \Longrightarrow^f_\Gamma \mathtt{s}$ *and neither* $\mathtt{s}$ *nor* $\mathtt{t}$ *contain any of the new constants* $\mathtt{c}_1, \mathtt{c}_2 \ldots$, *then* $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$.

Proof. f is defined in terms of an enumeration such that $\mathtt{r}$ always precedes $f(\mathtt{r})$. Thus if we repeatedly substituting each instance of $f(\mathtt{r})$ with $\mathtt{r}$ in a derivation, eventually all will be eliminated. But then instances of (con) depending on become trivial reductions $\mathtt{r} \Longrightarrow^f_\Gamma \mathtt{r}$ which can be removed without affecting the rest of the derivation. Certainly the first and final terms $\mathtt{t}$ and $\mathtt{s}$ are unaffected as they never contained any $f(\mathtt{r})$ in the first place. □

Definition 3.4. If $\mathtt{t}$ is a term let $w_\mathtt{t} = \{\mathtt{s} \mid \mathtt{t} \Longrightarrow^f_\Gamma \mathtt{s}\}$. Thus $w_\mathtt{t}$ is the closure of $\mathtt{t}$ under $\Longrightarrow^f_\Gamma$.

Definition 3.5. Define the **canonical λ-frame** $F_\lambda = \langle W_\lambda, \bullet_\lambda, \circ_\lambda, H_\lambda\rangle$:

$$W_\lambda = \{w_\mathtt{t} \mid \mathtt{t} \in \mathcal{L}^*_{\lambda_c}\} \qquad H_\lambda = \left\{\{w \mid \mathtt{t} \in w\} \mid w \in W_\lambda \text{ and } \mathtt{t} \in \mathcal{L}_{\lambda_c}\right\}$$

$$w_\mathtt{t} \bullet_\lambda w_\mathtt{s} = \{w \in W_\lambda \mid \mathtt{t}\cdot\mathtt{s} \in w\} \qquad w_\mathtt{t} \circ_\lambda w_\mathtt{s} = \{w \in W_\lambda \mid \mathtt{t} * \mathtt{s} \in w\}$$

Definition 3.6. Given $F_\lambda = \langle W_\lambda, \bullet_\lambda, \circ_\lambda, H_\lambda \rangle$, and a term $\mathtt{t}$ of $\mathcal{L}^*_\lambda$, let $\|\mathtt{t}\| = \{w \in W_\lambda \mid \mathtt{t} \in w\}$. Note that $H_\lambda = \left\{\|\mathtt{t}\| \mid \mathtt{t} \in \mathcal{L}_{\lambda_c}\right\}$

Remark 3.7. Given (*sub*) it is easy to see that $\bullet_\lambda \subseteq \circ_\lambda$. Frames where the converse does not hold are easy to construct (for example, Figure 3).

Theorem 3.8. *Let F_λ be the canonical intensional λ-frame (Definition 3.5), let $v(\mathtt{x}) = \|\mathtt{x}\|$ for any variable $\mathtt{x}$, and extend v so that $v(\mathtt{c}) = \mathtt{c}$ for any constant $\mathtt{c}$. Then for any term $\mathtt{t} \in \mathcal{L}_{\lambda_c}$, $[[\mathtt{t}]]^v = \|\mathtt{t}\|$.*

Proof. By induction on $\mathtt{t}$ we show that $w \in \|\mathtt{t}\|$ (i.e. $\mathtt{t} \in w$) iff $w \in [[\mathtt{t}]]^v$.

– $\mathtt{t}$ is a variable $\mathtt{x}$. Then $\|\mathtt{x}\| = v(\mathtt{x}) = [[\mathtt{x}]]^v$ by the definition of v.

– $\mathtt{t}$ is $\mathtt{t}_1{\cdot}\mathtt{t}_2$. Then $\mathtt{t}_1, \mathtt{t}_2 \in \mathcal{L}_{\lambda_c}$.
Suppose $\mathtt{t}_1{\cdot}\mathtt{t}_2 \in w$, and consider the worlds $w_{\mathtt{t}_1}$ and $w_{\mathtt{t}_2}$ in W_λ. If $\mathtt{s}_1 \in w_{\mathtt{t}_1}$ and $\mathtt{s}_2 \in w_{\mathtt{t}_2}$ then by Definition 3.4, $\mathtt{t}_1 \Longrightarrow^f_\Gamma \mathtt{s}_1$ and $\mathtt{t}_2 \Longrightarrow^f_\Gamma \mathtt{s}_2$. Thus $\mathtt{t}_1{\cdot}\mathtt{t}_2 \Longrightarrow^f_\Gamma \mathtt{s}_1{\cdot}\mathtt{s}_2$ and $\mathtt{s}_1{\cdot}\mathtt{s}_2 \in w$. Then by the definition of $\bullet_\lambda$ we have that $w \in w_{\mathtt{t}_1} \bullet_\lambda w_{\mathtt{t}_2}$. Furthermore, $w_{\mathtt{t}_1} \in \|\mathtt{t}_1\|$ and so by the induction hypothesis, $w_{\mathtt{t}_1} \in [[\mathtt{t}_1]]^v$. Similarly $w_{\mathtt{t}_2} \in [[\mathtt{t}_2]]^v$. Hence $w \in [[\mathtt{t}_1{\cdot}\mathtt{t}_2]]^v$ by Definition 2.13.

Conversely, suppose that $w \in [[\mathtt{t}_1{\cdot}\mathtt{t}_2]]^v$. Then there are $w_{\mathtt{s}_1}, w_{\mathtt{s}_2}$ such that $w_{\mathtt{s}_1} \in [[\mathtt{t}_1]]^v$ and $w_{\mathtt{s}_2} \in [[\mathtt{t}_2]]^v$ and $w \in w_{\mathtt{s}_1} \bullet_\lambda w_{\mathtt{s}_2}$. By the induction hypothesis $w_{\mathtt{s}_1} \in \|\mathtt{t}_1\|$ and $w_{\mathtt{s}_2} \in \|\mathtt{t}_2\|$. Then $\mathtt{s}_1 \Longrightarrow^f_\Gamma \mathtt{t}_1$ and $\mathtt{s}_2 \Longrightarrow^f_\Gamma \mathtt{t}_2$. Furthermore, by the construction of $\bullet_\lambda$, $\mathtt{s}_1{\cdot}\mathtt{s}_2 \in w$ and hence by (*cong*) $\mathtt{t}_1{\cdot}\mathtt{t}_2 \in w$.

– $\mathtt{t}$ is $\lambda\mathtt{x.s}$. $\mathtt{s} \in \mathcal{L}_{\lambda_c}$.
Suppose $\lambda\mathtt{x.s} \in w_1$. Suppose that $w_3 \in w_1 \circ_\lambda w_2$, and that $w_2 \in h$ for some $h \in H_\lambda$, then $h = \|\mathtt{r}\|$ for some term $\mathtt{r}$. By (ζ_f) we have that $\mathtt{r} \Longrightarrow^f_\Gamma \mathtt{c}$ and $\mathtt{c} \Longrightarrow^f_\Gamma \mathtt{r}$ for some $\mathtt{c} \in \mathcal{L}_{\lambda_c}$. So $h = \|\mathtt{c}\|$ and $\mathtt{c} \in w_2$. By the construction of $\circ_\lambda$, $\lambda\mathtt{x.s} * \mathtt{r} \in w_3$ and so $\mathtt{s[x/c]} \in w_3$ by (β^*), i.e. $w_3 \in \|\mathtt{s[x/c]}\|$. Since $\mathtt{s[x/c]} \in \mathcal{L}_{\lambda_c}$, it follows by the induction hypothesis that $\|\mathtt{s[x/c]}\| = [[\mathtt{s[x/c]}]]^v$. Furthermore by Lemma 2.16 $[[\mathtt{s[x/c]}]]^v = [[\mathtt{s}]]^{v[\mathtt{x}\mapsto[[\mathtt{c}]]^v]}$. But by the definition of v, $[[\mathtt{c}]]^v = \|\mathtt{c}\|$, and so $w_3 \in [[\mathtt{s}]]^{v[\mathtt{x}\mapsto\|\mathtt{c}\|]}$. But $h = \|\mathtt{c}\|$ so $w_3 \in [[\mathtt{s}]]^{v[\mathtt{x}\mapsto h]}$. Thus $w_1 \in \{w \mid \forall h \in H_\lambda. w \circ_\lambda h \subseteq [[\mathtt{s}]]^{v[\mathtt{x}\mapsto h]}\} = [[(\lambda\mathtt{x.s})]]^v$. Hence, $\|\lambda\mathtt{x.s}\| \subseteq [[(\lambda\mathtt{x.s})]]^v$

Conversely, suppose that $\lambda\mathtt{x.s} \notin w_\mathtt{r}$ for some $\mathtt{r}$. Let $\mathtt{y}$ be a variable not free in $\mathtt{r}$ or $\mathtt{s}$ and consider the worlds $w_\mathtt{y}$ and $w_{\mathtt{r}*\mathtt{y}}$. If $\mathtt{s[x/y]} \in w_{\mathtt{r}*\mathtt{y}}$ then $\mathtt{r} * \mathtt{y} \Longrightarrow^f_\Gamma \mathtt{s[x/y]}$, so $\lambda\mathtt{y.(r} * \mathtt{y)} \Longrightarrow^f_\Gamma \lambda\mathtt{y(s[x/y])}$ by (ξ). But by our choice of $\mathtt{y}$, (η) entails that $\mathtt{r} \Longrightarrow^f_\Gamma \lambda\mathtt{y.(r} * \mathtt{y)}$. So $\mathtt{r} \Longrightarrow^f_\Gamma \lambda\mathtt{y.s[x/y]}$, which contradicts our initial supposition that $\lambda\mathtt{x.s} \notin w_\mathtt{r}$, therefore $\mathtt{s[x/y]} \notin w_{\mathtt{r}*\mathtt{y}}$. In other words $w_{\mathtt{r}*\mathtt{y}} \notin \|\mathtt{s[x/y]}\|$. But $\mathtt{s[x/y]} \in \mathcal{L}_{\lambda_c}$, so by the induction hypothesis $w_{\mathtt{r}\cdot\mathtt{y}} \notin [[\mathtt{s[x/y]}]]^v$. Since $[[\mathtt{y}]]^v = \|\mathtt{y}\|$, it follows by Lemma 2.16 that $w_{\mathtt{r}\cdot\mathtt{y}} \notin [[\mathtt{s}]]^{v[\mathtt{x}\mapsto\|\mathtt{y}\|]}$. But clearly $w_{\mathtt{r}*\mathtt{y}} \in w_\mathtt{r} \circ_\lambda w_\mathtt{y}$, so it follows that $w_{\mathtt{r}} \notin \{w \mid \forall h \in H_\lambda . w \circ_\lambda h \subseteq [\![\mathtt{s}]\!]^{v[\mathtt{x}\mapsto h]}\}$. By the semantics of $\lambda\mathtt{y}.\mathtt{s}$ this means that $w_{\mathtt{r}} \notin [\![(\lambda\mathtt{y}.\mathtt{s})]\!]^v$. Hence, since every $w \in W_\lambda$ is $w_{\mathtt{r}}$ for some $\mathtt{r}$, $[\![(\lambda\mathtt{x}.\mathtt{s})]\!]^v \subseteq \|\lambda\mathtt{x}.\mathtt{s}\|$. □

Lemma 3.9. *If v_1, v_2 are any valuations on a frame F that such that*

1. *$v_1(\mathtt{x}) = v_2(\mathtt{x})$ for any variable $\mathtt{x}$ that occurs free in $\mathtt{t}$,*
2. *v_1, v_2 are extended so that $v_1(\mathtt{c}) = v_2(\mathtt{c})$ for any constant $\mathtt{c}$ that occurs in $\mathtt{t}$,*

then $[\![\mathtt{t}]\!]^{v_1} = [\![\mathtt{t}]\!]^{v_2}$.

Proof. By an easy induction on $\mathtt{t}$. □

Lemma 3.10. *If there is a valuation v on a frame F such that $\{[\![\mathtt{t}]\!]^v \mid \mathtt{t} \in \mathcal{L}_\lambda\} = H$, then F is faithful. Hence the canonical frame F_λ is faithful.*

Proof. Suppose there is such a v, then we must show that for any valuation v' and any term $\mathtt{t} \in \mathcal{L}_\lambda$ that $[\![\mathtt{t}]\!]^{v'} \in H$. By the definition of a valuation, $[\![\mathtt{x}]\!]^{v'} \in H$ for any variable $\mathtt{x}$. So if $[\![\mathtt{t}]\!]^{v'} \notin H$ then by Lemma 3.9

$$[\![\mathtt{t}]\!]^{v[\mathtt{x}_1 \mapsto [\![\mathtt{x}_1]\!]^{v'} \ldots \mathtt{x}_n \mapsto [\![\mathtt{x}_n]\!]^{v'}]} \notin H$$

where $\mathtt{x}_1 \ldots \mathtt{x}_n$ are the free variables of $\mathtt{t}$. Now, by assumption, v is such that every $h \in H$ is $[\![\mathtt{s}]\!]^v$ for some $\mathtt{s}$. It follows then that we can choose $\mathtt{s}_1 \ldots \mathtt{s}_n$ such that $[\![\mathtt{s}_i]\!]^v = v[\mathtt{x}_i \mapsto [\![\mathtt{x}_i]\!]^{v'}]$, and so:

$$[\![\mathtt{t}]\!]^{v[\mathtt{x}_1 \mapsto [\![\mathtt{s}_1]\!]^{v} \ldots \mathtt{x}_n \mapsto [\![\mathtt{s}_n]\!]^{v}]} \notin H$$

This entails, by Theorem 2.16 that $[\![\mathtt{t}[\mathtt{x}_i/\mathtt{s}_i]]\!]^v \notin H$. But this contradicts the assumption that $\{[\![\mathtt{t}]\!]^v \mid \mathtt{t} \in \mathcal{L}_\lambda\} = H$. □

Lemma 3.11. *F_λ is Γ-sensitive.*

Proof. We must argue that for $\langle \mathtt{t}_1, \mathtt{t}_2 \rangle \in \Gamma$ and any v, $[\![\mathtt{t}_1]\!]^v \subseteq [\![\mathtt{t}_2]\!]^v$. Let $\mathtt{x}_1 \ldots \mathtt{x}_n$ be the free variables of $\mathtt{t}_1$ and $\mathtt{t}_2$. Then $v(\mathtt{x}_i)$ is some $\|\mathtt{s}_i\| \in H_\lambda$.

Let v' be a valuation extended such that $v'(\mathtt{r}) = \|\mathtt{r}\|$ for for any variable or constant $\mathtt{r}$ (i.e. v' meets the condition of Theorem 3.8). Then:

$$\begin{aligned}
[\![\mathtt{t}_1]\!]^v &= [\![\mathtt{t}_1]\!]^{v'[\mathtt{x}_1 \mapsto [\![\mathtt{x}_1]\!]^{v} \ldots \mathtt{x}_n \mapsto [\![\mathtt{x}_n]\!]^{v}]} && \text{Lemma 3.9} \\
&= [\![\mathtt{t}_1]\!]^{v'[\mathtt{x}_1 \mapsto \|\mathtt{s}_1\| \ldots \mathtt{x}_n \mapsto \|\mathtt{s}_n\|]} && \\
&= [\![\mathtt{t}_1]\!]^{v'[\mathtt{x}_1 \mapsto [\![\mathtt{s}_1]\!]^{v'} \ldots \mathtt{x}_n \mapsto [\![\mathtt{s}_n]\!]^{v'}]} && \text{Theorem 3.8} \\
&= [\![\mathtt{t}_1[\mathtt{x}_i/\mathtt{s}_i]]\!]^{v'} && \text{Lemma 2.16} \\
&= \|\mathtt{t}_1[\mathtt{x}_i/\mathtt{s}_i]\| && \text{Theorem 3.8}
\end{aligned}$$

and similarly for $\mathtt{t}_2$. But $\mathtt{t}_1[\mathtt{x}_i/\mathtt{s}_i] \Longrightarrow_\Gamma \mathtt{t}_2[\mathtt{x}_i/\mathtt{s}_i]$ by (*ass*), and so $\|\mathtt{t}_1[\mathtt{x}_i/\mathtt{s}_i]\| \subseteq \|\mathtt{t}_2[\mathtt{x}_i/\mathtt{s}_i]\|$ and so $[\![\mathtt{t}_1]\!]^v \subseteq [\![\mathtt{t}_2]\!]^v$. □

Theorem 3.12. $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ *if and only if* $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ *for all* Γ*-sensitive models.*[4]

Proof. The left-right direction is Theorem 2.25.

If $\mathtt{t} \not\Longrightarrow_\Gamma \mathtt{s}$ then $\mathtt{s} \notin w_{\mathtt{t}}$ in F_λ. Therefore $\|\mathtt{t}\| \nsubseteq \|\mathtt{s}\|$ and so by Theorem 3.8 there is a valuation v such that $[[\mathtt{t}]]^v \nsubseteq [[\mathtt{s}]]^v$ on the canonical frame F_λ. Furthermore, by Lemmas 3.10 and 3.11, F_λ is faithful and Γ-sensitive. □

Corollary 3.13. *If* $\mathtt{t}$ *and* $\mathtt{s}$ *are terms of* $\mathcal{L}_\lambda$ *then* $\mathtt{t} \longrightarrow_\Gamma \mathtt{s}$ *if and only if* $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ *for all* Γ*-sensitive models.*

Proof. Using Theorem 2.8 and the assumption that $\mathtt{t}$ and $\mathtt{s}$ are terms of $\mathcal{L}_\lambda$ we get that $\mathtt{t} \longrightarrow_\Gamma \mathtt{s}$ if and only if $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ □

Definition 3.14. An **extensional frame** is an intensional frame where $\bullet = \circ$, we may define them simply as a triple $\langle W, \bullet, H\rangle$. Similarly an **extensional model** is a pair $\langle F, v\rangle$ where F is an extensional frame.

Corollary 3.15. *Let* $\Gamma = \{\langle \mathtt{x}, \lambda\mathtt{y}.(\mathtt{x}{\cdot}\mathtt{y})\rangle\}$. *Then* $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ *if and only if* $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ *for any faithful extensional model.*

Proof. For the left-right direction it is a simple matter to apply the reasoning of Theorem 2.25. For the right-left direction it is enough to note that:

$$\mathtt{t}[\mathtt{x}{:-}\mathtt{s} * \mathtt{r}] \overset{(ass)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\lambda\mathtt{y}(\mathtt{s}{\cdot}\mathtt{y}) * \mathtt{r}] \overset{(\beta^*)}{\Longrightarrow}_\Gamma \mathtt{t}[\mathtt{x}{:-}\mathtt{s}{\cdot}\mathtt{r}]$$

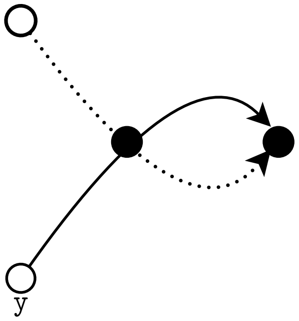
so in the construction of the canonical frame F_λ of Theorem 3.8, $\bullet_\lambda = \circ_\lambda$. □

Remark 3.16. An extensional frame satisfies η-expansion. An intensional frame is like an extensional frame except with an additional 'outer' application function $\circ$. We interpret λ in terms of the outer function and application in terms of the inner function $\bullet$ to block η-expansion (Definition 2.13). η-expansion will prove useful in constructing models of λ-equality in Section 4. Other authors have also noted reasons to include η-expansion in models [13].

Remark 3.17. Given 3.15, we can say that λ-reduction with η-expansion is *complete* for extensional frames.

Remark 3.18. Notice also a crucial purpose served by H in the completeness proof. Any subset of a frame is a potential denotation of a λ-term, and H may be seen as listing the subsets that actually are denotations of λ-terms. We have used this distinction to characterise intensional λ-reduction.

[4] Equivalently: $\mathtt{t} \Longrightarrow_\Gamma \mathtt{s}$ if and only if $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ for any valuation v on any Γ-sensitive frame.



A solid arrow passing from w_1 through w_2 to w_3 represents that $w_3 \in w_1 \bullet w_2$, and a dotted arrow represents that $w_3 \in w_1 \circ w_2$. H can be set as $\mathcal{P}(W)$ and $v(\mathtt{y})$ is as indicated. Then the two worlds on the left (unfilled) both are in $[\![\lambda\mathtt{x}.(\mathtt{y}\cdot\mathtt{x})]\!]^v$ and $[\![\lambda\mathtt{x}.(\mathtt{y} * \mathtt{x})]\!]^v$, but only one is in $[\![\mathtt{y}]\!]^v$.

Figure 3: A counterexample to η-reduction in an intensional model where W contains 4 worlds.

We took an (intensional) set of λ-reductions Γ (in $\mathcal{L}_\lambda$) and we extended it using $*$ to help us interpret λ (Definition 2.5 and Theorem 2.8). Then, when we constructed the frame (Definition 3.5) for Γ we left out of H the denotations depending explicitly on $*$. We obtained a frame which is sensitive to all the reductions of Γ, in the original language $\mathcal{L}_\lambda$, but where the interpretation of λ still depends on $*$ which is not mentioned in Γ (Theorem 3.8).

As we shall see, this provides a simple characterisation of intensional and extensional λ-abstraction. Abusing notation somewhat: extensional $\lambda\mathtt{x}.\mathtt{t}$ is something that maps objects h in the domain to $\mathtt{t}(h)$; intensional $\lambda\mathtt{x}.\mathtt{t}$ is something maps objects h in the domain *and also some in a hidden domain* to $\mathtt{t}(h)$. Furthermore, the 'hidden' objects are the denotations of terms in $\mathcal{L}^*_\lambda$ that require $*$.

3.1 η-reduction

As already noted, if $\mathtt{x}$ is not free in $\mathtt{t}$, then $[\![\mathtt{t}]\!]^v \subseteq [\![\lambda\mathtt{x}.(\mathtt{t} * \mathtt{x})]\!]^v$ in any intensional frame. That is, η-expansion is satisfied by any frame, but what about η-reduction? Figure 3 gives an example of a simple frame where $[\![\mathtt{y}]\!]^v \not\subseteq [\![\lambda\mathtt{x}.(\mathtt{y}\cdot\mathtt{x})]\!]^v$ (and since $\bullet \subseteq \circ$, also $[\![\mathtt{y}]\!]^v \not\subseteq [\![\lambda\mathtt{x}.(\mathtt{y} * \mathtt{x})]\!]^v$).

We can characterise η-reduction syntactically easily enough:

Definition 3.19. Let $\boldsymbol{\eta}^- = \{\langle\lambda\mathtt{x}.(\mathtt{y}\cdot\mathtt{x}), \mathtt{y}\rangle\}$.

Then $\mathtt{t} \Longrightarrow_{\boldsymbol{\eta}^-} \mathtt{s}$ is the relation we want. Furthermore, we can use the completeness theorem 3.12 to describe a class of models for which this relation is complete:

Definition 3.20. A frame is η**-reductive** when $\bigcap_{h' \in H}\{w \mid w \circ h' \subseteq h \bullet h'\} \subseteq h$ for any h.

Theorem 3.21. $\mathtt{t} \Longrightarrow_{\boldsymbol{\eta}^-} \mathtt{s}$ *iff* $\mathtt{t} \subseteq \mathtt{s}$ *in all* η*-reductive models.*

Proof. It is straightforward to verify that $\mathtt{t} \Longrightarrow_{\boldsymbol{\eta}^-} \mathtt{s}$ implies that $[[\mathtt{t}]]^v \subseteq [[\mathtt{s}]]^v$ in all η-reductive models. Conversely, if $\mathtt{t} \not\Longrightarrow_{\boldsymbol{\eta}^-} \mathtt{s}$ then $[[\mathtt{t}]]^v \nsubseteq [[\mathtt{s}]]^v$ in the canonical model for $\boldsymbol{\eta}^-$:

$$\bigcap_{h' \in H_\lambda} \{w \mid w \circ h' \subseteq \|\mathtt{t}\| \bullet h'\} = \|\lambda\mathtt{x}.(\mathtt{t}\cdot\mathtt{x})\| \subseteq \|\mathtt{t}\|$$

since each $h \in H$ is $\|\mathtt{t}\|$ for some $\mathtt{t}$. □

4 Equality

Definition 4.1. Let $\boldsymbol{\beta} = \{\langle \mathtt{t}, \lambda\mathtt{x}.\mathtt{t}\cdot\mathtt{x}] \rangle \mid \mathtt{t} \in \mathcal{L}_\lambda\}$.

Corollary 4.2. *When restricted to* $\mathcal{L}_\lambda$*,* $\Longrightarrow_{\boldsymbol{\beta}}$ *is the familiar relation of (intensional)* λ*-equality, and by Theorem 3.12 is complete for* $\boldsymbol{\beta}$*-sensitive models.*

Remark 4.3. Corollary 4.2 is itself not so significant as it only tells us half the story about what these models look like, and does not tell us if there are any non-trivial ones. Of course, given independent nontriviality proofs for λ-equality,[5] we can use Theorem 3.12 to conclude that there are nontrivial $\boldsymbol{\beta}$-sensitive models. This section is concerned with producing a purely semantic characterisation of $\boldsymbol{\beta}$-sensitivity. In fact, in characterising $\boldsymbol{\beta}$-sensitivity, we can complete Definition 2.17 and provide a semantic characterisation of faithfulness.

The strategy we shall employ is as follows. First we introduce some shorthands to stand in for constructions involving λ-expressions, so for example $\mathbf{K}\cdot\mathtt{z}$ will stand in for $(\lambda\mathtt{x}\lambda\mathtt{y}.\mathtt{x})\cdot\mathtt{z}$. This will allow us to work with certain complex λ-expressions as if they are free of the symbol λ. Then, for each $\mathtt{t}$ we describe a new term $[\mathtt{x}]\mathtt{t}$, constructible only out of application and the new shorthands (effectively the familiar combinator abstraction of [12, p.26], but extended to a language that includes the λ-operator). Then, with the help of the completeness theorem 3.12, we describe conditions in which a model (or frame) entails that $[[\mathtt{t}[\mathtt{x}/\mathtt{s}]]]^v = [[[\mathtt{x}]\mathtt{t}\cdot\mathtt{s}]]^v$. It then turns out that $[[[\mathtt{x}]\mathtt{t}]]^v \subseteq [[\lambda\mathtt{x}.\mathtt{t}]]^v$ and we thereby obtain models of β-expansion.

[5] Nontriviality follows syntactically from the Church-Rosser property [12, Ch. A2], the cut-elimination theorem of [5]; and it follows semantically from Scott's famous model D_ω [12, Ch. 16], among others.

Definition 4.4. Define the following shorthands:

$$\begin{aligned}\mathbf{K} &= \lambda \mathtt{xy.x}\\ \mathbf{C} &= \lambda \mathtt{xyz.((x{\cdot}z){\cdot}y))}\\ \mathbf{S} &= \lambda \mathtt{xyz.((x{\cdot}z){\cdot}(y{\cdot}z))}\end{aligned}$$

Definition 4.5. – Say that an instance of λ in a term t is **free** if it is not part of an occurrence of **K**, **C**, **S** in t.

– When defining or proving a property of a term t, we write '**by induction on (l, d)**' to describe an induction on the pair (l, d), lexicographically ordered, where d is the number of occurrences of $\cdot$ in t and l is the number of occurrences of λ in t that are not free.[6]

Definition 4.6. For any $\mathtt{t} \in \mathcal{L}_\lambda$, define [x]t by induction on (l, d).

1. (a) [x]x is $(\mathbf{S}{\cdot}\mathbf{K}){\cdot}\mathbf{K}$

   (b) [x]r is $\mathbf{K}{\cdot}\mathtt{r}$ if r is **K**, **C**, **S**, or any variable distinct from x.

2. [x](s·r) is $(\mathbf{S}{\cdot}$[x]s)·[x]r

3. – [x]λz.s is $\mathbf{C}{\cdot}$[z][x]s

   – [z]λz.s is $\mathbf{C}{\cdot}$[z][x]s, where x is distinct from z and does not occur in s.

Lemma 4.7. *If* $\mathtt{t} \in \mathcal{L}_\lambda$ *then* $[\mathtt{x}]\mathtt{t} \in \mathcal{L}^*_\lambda$ *and (1) is well defined, (2) contains no free instances of* λ*, and (3) contains no free occurrences of* x.

Proof. By induction on (l, d).

– If t is atomic or of the form s·r or s $*$ r then the result is easily proved.

– If t is λz.s then by the induction hypothesis [x]s is well defined and contains no free occurrences of λ. So the induction hypothesis applies again and the same may be said of [z][x]s. It then follows easily that the properties 1, 2 and 3 hold for [x]λz.s.

□

Lemma 4.8. *Suppose* $\mathtt{t} \in \mathcal{L}_\lambda$ *and let variable* v *not occur in* t*, then* [x](t[y/v]) = ([x]t)[y/v]

Proof. By induction on (l, d).

[6] More loosely, if we were to treat **K**, **C**, **S** as constants in t (i.e. not containing λ at all), then l would be the number of occurrences of λ in t.

– If $\mathtt{t}$ is $\mathtt{x}$ then $\mathtt{x}[\mathtt{y}/\mathtt{v}] = \mathtt{x}$ and so

$$\begin{array}{rcll} [\mathtt{x}]\mathtt{x} & = & (\mathbf{S}\cdot\mathbf{K}\cdot)\mathbf{K} & \text{Def. 4.6} \\ & = & ([\mathtt{x}]\mathtt{x})[\mathtt{y}/\mathtt{v}] & \mathtt{y} \notin (\mathbf{S}\cdot\mathbf{K}\cdot)\mathbf{K} \end{array}$$

– If $\mathtt{t}$ is $\mathtt{y}$ then $\mathtt{y}[\mathtt{y}/\mathtt{v}] = \mathtt{v}$ and $[\mathtt{x}]\mathtt{y} = \mathbf{K}\cdot\mathtt{y}$. So:

$$\begin{array}{rcll} [\mathtt{x}]\mathtt{v} & = & \mathbf{K}\cdot\mathtt{v} & \text{Def. 4.6} \\ & = & (\mathbf{K}\cdot\mathtt{y})[\mathtt{y}/\mathtt{v}] & \mathtt{y} \notin \mathbf{K} \\ & = & ([\mathtt{x}]\mathtt{y})[\mathtt{y}/\mathtt{v}] & \end{array}$$

– The case where $\mathtt{t}$ is $\mathbf{K}$, $\mathbf{C}$, $\mathbf{S}$ or some variable other than $\mathtt{x}$ or $\mathtt{y}$ is similar.

– If $\mathtt{t}$ is $\mathtt{s}\cdot\mathtt{r}$ or $\mathtt{s} * \mathtt{r}$ then the result follows easily by the induction hypothesis.

– If $\mathtt{t}$ is $\lambda\mathtt{z}.\mathtt{s}$, then we may assume that $\mathtt{z}$ is not $\mathtt{x}$, then

$$\begin{array}{rcll} [\mathtt{x}](\mathtt{t}[\mathtt{y}/\mathtt{v}]) & = & \mathbf{C}\cdot[\mathtt{z}][\mathtt{x}](\mathtt{s}[\mathtt{y}/\mathtt{v}]) & \text{Def. 4.6} \\ & = & \mathbf{C}\cdot([\mathtt{z}][\mathtt{x}]\mathtt{s})[\mathtt{y}/\mathtt{v}] & \text{ind. hyp, Lemma 4.7} \\ & = & (\mathbf{C}\cdot[\mathtt{z}][\mathtt{x}]\mathtt{s})[\mathtt{y}/\mathtt{v}] & \mathtt{y} \notin \mathbf{C} \\ & = & ([\mathtt{x}]\mathtt{t})[\mathtt{y}/\mathtt{v}] & \end{array}$$

□

Theorem 4.9. *If* $\mathtt{t} \in \mathcal{L}_\lambda$, *then for any* $M = \langle F, v\rangle$, $[\![[\mathtt{x}]\mathtt{t} * \mathtt{s}]\!]^v \subseteq [\![\mathtt{t}[\mathtt{x}/\mathtt{s}]]\!]^v$

Proof. By induction on (l, d).

We appeal to known facts about β-reduction and Theorem 3.12 (completeness).

– $\mathtt{t} = \mathtt{x}$. Then $[\mathtt{x}]\mathtt{t} * \mathtt{s}$ is $((\mathbf{S}\cdot\mathbf{K}\cdot)\mathbf{K})\cdot\mathtt{s}$, and it is easy to show that that

$$((\mathbf{S}\cdot\mathbf{K}\cdot)\mathbf{K}) * \mathtt{s} \Longrightarrow_{\varnothing} \mathtt{s}$$

So the result follows by Theorem 3.12.

– The argument is similar for the case where $\mathtt{t}$ is a variable $\mathtt{y} \neq \mathtt{x}$ or $\mathbf{K}$, $\mathbf{C}$, $\mathbf{S}$. We appeal to the easily shown fact that:

$$(\mathbf{K}\cdot\mathtt{t}) * \mathtt{s} \Longrightarrow_{\varnothing} \mathtt{t}$$

– $\mathtt{t} = \mathtt{t}_1\cdot\mathtt{t}_2$. Then

$$\begin{array}{rcl} ((\mathbf{S}\cdot[\mathtt{x}]\mathtt{t}_1) * [\mathtt{x}]\mathtt{t}_2)\cdot\mathtt{s} & \Longrightarrow_{\varnothing} & \mathtt{t}_1[\mathtt{x}/\mathtt{s}]\cdot\mathtt{t}_2[\mathtt{x}/\mathtt{s}] \\ & = & (\mathtt{t}_1\cdot\mathtt{t}_2)[\mathtt{x}/\mathtt{s}] \end{array}$$

and the result follows as above.

– The argument is similar for the case where $\mathtt{t} = \mathtt{t}_1 * \mathtt{t}_2$

– If $\mathtt{t}$ is $\lambda\mathtt{y}.\mathtt{r}$, then choose a variable $\mathtt{z}$ that does not occur in $\mathtt{r}$ or $\mathtt{s}$. Now, $\mathtt{s}$ and $[\mathtt{x}]\mathtt{s}$ contain fewer free instances of λ than $\mathtt{t}$ (Lemma 4.7), so given Theorem 3.12 we may apply the induction hypothesis as follows:

$$\begin{array}{rcll}
\llbracket(\mathbf{C}\cdot[\mathtt{y}][\mathtt{x}]\mathtt{r})\cdot\mathtt{s}\rrbracket^v & \subseteq & \llbracket\lambda\mathtt{z}.\big(((\mathbf{C}\cdot[\mathtt{y}][\mathtt{x}]\mathtt{r})\cdot\mathtt{s}) * \mathtt{z}\big)\rrbracket^v & \text{Thrm 2.25}\\
 & \subseteq & \llbracket\lambda\mathtt{z}.\big(([\mathtt{y}][\mathtt{x}]\mathtt{r}\cdot\mathtt{z})\cdot\mathtt{s}\big)\rrbracket^v & \text{Thrm 3.12}\\
 & \subseteq & \llbracket\lambda\mathtt{z}.([\mathtt{x}]\mathtt{r}[\mathtt{y}/\mathtt{z}]\cdot\mathtt{s})\rrbracket^v & \text{Ind. Hyp.}\\
 & \subseteq & \llbracket\lambda\mathtt{z}.(\mathtt{t}[\mathtt{y}/\mathtt{z},\mathtt{x}/\mathtt{s}])\rrbracket^v & \text{Ind. Hyp.}\\
 & = & \llbracket\lambda\mathtt{y}.(\mathtt{t}[\mathtt{x}/\mathtt{s}])\rrbracket^v &
\end{array}$$

□

Definition 4.10. A frame is **λ-complete** when for any $h_1, h_2, h_3 \in H$

1. $h_1 = (\llbracket\mathbf{K}\rrbracket^v \bullet h_1) \bullet h_2$
2. $(h_1 \bullet h_3) \bullet (h_2 \bullet h_3) = ((\llbracket\mathbf{S}\rrbracket^v \bullet h_1) \bullet h_2) \bullet h_3$
3. $\bigcap_{h \in H}\{w \mid (h_1 \bullet h) \bullet h_2\} = (\llbracket\mathbf{C}\rrbracket^v \bullet h_1) \bullet h_2$

A model $\langle F, v\rangle$ is λ-complete if F is.

Remark 4.11. Notice that no $h \in H$ can be empty if F is a non-trivial λ-complete frame. For if $\varnothing \in H$ then for any $h \in H$, $h = (\llbracket\mathbf{K}\rrbracket^v \bullet h) \bullet \varnothing = \varnothing$, so then $H = \{\varnothing\}$ and $\llbracket\mathtt{t}\rrbracket^v = \llbracket\mathtt{s}\rrbracket^v = \varnothing$ for any $\mathtt{t}, \mathtt{s}$ and v.

We could have equivalently defined λ-complete frames by requiring that, for any v $\llbracket\mathtt{x}\rrbracket^v = \llbracket(\mathbf{K}\cdot\mathtt{x})\cdot\mathtt{y}\rrbracket^v$, $\llbracket(\mathtt{x}\cdot\mathtt{z})\cdot(\mathtt{y}\cdot\mathtt{z})\rrbracket^v = \llbracket((\mathbf{S}\cdot\mathtt{x})\cdot\mathtt{y})\cdot\mathtt{z}\rrbracket^v$, $\llbracket\lambda\mathtt{z}.((\mathtt{x}\cdot\mathtt{z})\cdot\mathtt{y})\rrbracket^v = \llbracket(\mathbf{C}\cdot\mathtt{x})\cdot\mathtt{y}\rrbracket^v$ and so on. But 4.10 is preferable as its form is less dependent on the syntax. Since $\mathbf{K}$, $\mathbf{S}$ and $\mathbf{C}$ are closed terms, we could even go further and replace $\llbracket\mathbf{K}\rrbracket^v$, $\llbracket\mathbf{S}\rrbracket^v$ and $\llbracket\mathbf{C}\rrbracket^v$ with purely semantic expressions using Definition 2.13.

Theorem 4.12. *For any λ-complete $\langle F, v\rangle$, if $\mathtt{t} \in \mathcal{L}_\lambda$ then $\llbracket\mathtt{t}[\mathtt{x}/\mathtt{s}]\rrbracket^v \subseteq \llbracket[\mathtt{x}]\mathtt{t}\cdot\mathtt{s}\rrbracket^v$.*

Proof. Again, we proceed by induction on (l, d).

– $\mathtt{t} = \mathtt{x}$. Then $\mathtt{x}[\mathtt{x}/\mathtt{s}] = \mathtt{s}$ and it is not hard to see that the definition of lambda completeness (4.10) implies that $\llbracket\mathtt{s}\rrbracket^v = \llbracket((\mathbf{S}\cdot\mathbf{K})\cdot\mathbf{K})\cdot\mathtt{s}\rrbracket^v$.

– The argument is similar for the case where $\mathtt{t} = \mathtt{y} \neq \mathtt{x}$ or $\mathtt{t}$ is $\mathbf{K}$, $\mathbf{C}$, $\mathbf{S}$.

– $\mathtt{t} = \mathtt{t}_1\cdot\mathtt{t}_2$. Then:

$$\begin{array}{rcll}
\llbracket(\mathtt{t}_1\cdot\mathtt{t}_2)[\mathtt{x}/\mathtt{s}]\rrbracket^v & = & \llbracket\mathtt{t}_1[\mathtt{x}/\mathtt{s}]\rrbracket^v \bullet \llbracket\mathtt{t}_2[\mathtt{x}/\mathtt{s}]\rrbracket^v & \\
 & \subseteq & \llbracket([\mathtt{x}]\mathtt{t}_1\cdot\mathtt{s})\rrbracket^v \bullet \llbracket([\mathtt{x}]\mathtt{t}_2\cdot\mathtt{s})\rrbracket^v & \text{Ind. Hyp.}\\
 & = & (\llbracket[\mathtt{x}]\mathtt{t}_1\rrbracket^v \bullet \llbracket\mathtt{s}\rrbracket^v) \bullet (\llbracket[\mathtt{x}]\mathtt{t}_2\rrbracket^v \bullet \llbracket\mathtt{s}\rrbracket^v) & \\
 & = & ((\llbracket\mathbf{S}\rrbracket^v \bullet \llbracket[\mathtt{x}]\mathtt{t}_1\rrbracket^v) \bullet \llbracket[\mathtt{x}]\mathtt{t}_2\rrbracket^v) \bullet \llbracket\mathtt{s}\rrbracket^v & \text{Def 4.10}\\
 & = & \llbracket((\mathbf{S}\cdot[\mathtt{x}]\mathtt{t}_1)\cdot[\mathtt{x}]\mathtt{t}_2)\cdot\llbracket\mathtt{s}\rrbracket^v\rrbracket^v & \text{Def. 2.13}\\
 & = & \llbracket([\mathtt{x}]\mathtt{t}_1\cdot\mathtt{t}_2)\cdot\mathtt{s}\rrbracket^v & \text{Def. 4.6}
\end{array}$$

and the result follows as above.

– Suppose $\mathtt{t}$ is $\lambda\mathtt{y}.\mathtt{r}$. Let $\mathtt{z}$ be chosen so that it does not occur in $\mathtt{t}$ or $\mathtt{s}$. Then using Lemma 2.24:

$$\begin{array}{rcll}
\llbracket \lambda\mathtt{y}.(\mathtt{r}[\mathtt{x}/\mathtt{s}])\rrbracket^v & = & \llbracket \lambda\mathtt{z}.(\mathtt{r}[\mathtt{y}/\mathtt{z},\mathtt{x}/\mathtt{s}])\rrbracket^v & \\
 & \subseteq & \llbracket \lambda\mathtt{z}.\big([\mathtt{x}](\mathtt{r}[\mathtt{y}/\mathtt{z}])\cdot\mathtt{s}\big)\rrbracket^v & \text{Ind. Hyp.} \\
 & = & \llbracket \lambda\mathtt{z}.\big(([\mathtt{x}]\mathtt{r})[\mathtt{y}/\mathtt{z}]\cdot\mathtt{s}\big)\rrbracket^v & \text{Lemma 4.8} \\
 & \subseteq & \llbracket \lambda\mathtt{z}.\big(([\mathtt{y}][\mathtt{x}]\mathtt{r}\cdot\mathtt{z})\cdot\mathtt{s}\big)\rrbracket^v & \text{Ind. Hyp.} \\
 & \subseteq & \llbracket (\mathbf{C}\cdot[\mathtt{y}][\mathtt{x}]\mathtt{r})\cdot\mathtt{s}\rrbracket^v & \text{Def. 4.10} \\
 & = & \llbracket ([\mathtt{x}]\lambda\mathtt{y}.\mathtt{r})\cdot\mathtt{s}\rrbracket^v & \text{Def. 4.6}
\end{array}$$

□

Corollary 4.13. *If* $\mathtt{t} \in \mathcal{L}_\lambda$ *then* $\llbracket \mathtt{t}[\mathtt{x}/\mathtt{s}]\rrbracket^v = \llbracket \lambda.\mathtt{xt}\cdot\mathtt{s}\rrbracket^v$ *for any λ-complete model* $\langle F, v\rangle$.

Proof. Since $\bullet \subseteq \circ$ we have that $\llbracket \lambda\mathtt{x}.\mathtt{t}\cdot\mathtt{s}\rrbracket^v \subseteq \llbracket \lambda\mathtt{x}.\mathtt{t} \ast \mathtt{s}\rrbracket^v$, and so Theorem 4.9 entails that $\llbracket \lambda\mathtt{x}.\mathtt{t}\cdot\mathtt{s}\rrbracket^v \subseteq \llbracket \mathtt{t}[\mathtt{x}/\mathtt{s}]\rrbracket^v$.

And conversely:

$$\begin{array}{rcll}
\llbracket \mathtt{t}[\mathtt{x}/\mathtt{s}]\rrbracket^v & \subseteq & \llbracket [\mathtt{x}]\mathtt{t}\cdot\mathtt{s}\rrbracket^v & \text{Thrm 4.12} \\
 & \subseteq & \llbracket \lambda\mathtt{x}.([\mathtt{x}]\mathtt{t} \ast \mathtt{x})\cdot\mathtt{s}\rrbracket^v & \text{Thrm 3.12, Lemma 2.24, Lemma 4.7} \\
 & \subseteq & \llbracket \lambda\mathtt{x}.\mathtt{t}\cdot\mathtt{s}\rrbracket^v & \text{Thrm 4.9}
\end{array}$$

□

We now have a means of characterising β-equality semantically.

Corollary 4.14. *If* $\mathtt{t}, \mathtt{s} \in \mathcal{L}_\lambda$ *then,* $\mathtt{t} \Longrightarrow_{\boldsymbol{\beta}} \mathtt{s}$ *iff* $\llbracket \mathtt{t}\rrbracket^v \subseteq \llbracket \mathtt{s}\rrbracket^v$ *for all λ-complete models* $\langle F, v\rangle$.

Proof. By 4.13 if $\langle \mathtt{t}, \mathtt{s}\rangle \in \boldsymbol{\beta}$ then $\llbracket \mathtt{t}\rrbracket^v \subseteq \llbracket \mathtt{s}\rrbracket^v$ in all λ-complete frames. Furthermore if $\mathtt{t} \not\Longrightarrow_{\boldsymbol{\beta}} \mathtt{s}$ then $\llbracket \mathtt{t}\rrbracket^v \nsubseteq \llbracket \mathtt{s}\rrbracket^v$ in the canonical model for $\boldsymbol{\beta}$. It is not hard to verify that the canonical frame is λ-complete. □

Corollary 4.15. *A frame is λ-complete iff it is* $\boldsymbol{\beta}$*-sensitive*

Definition 4.16. A frame is **fully extensional** when $h = \bigcap_{h' \in H}\{w \mid w \circ h' \subseteq h \bullet h'\}$ for all $h \in H$. A model $\langle F, v\rangle$ is fully extensional when F is.

Remark 4.17. Looking at Definition 2.13 $h = \llbracket \mathtt{t}\rrbracket^v$ implies $\bigcap_{h' \in H}\{w \mid w \circ h' \subseteq h \bullet h'\} = \llbracket \lambda\mathtt{x}.(\mathtt{t}\cdot\mathtt{x})\rrbracket^v$ for $\mathtt{x} \notin \mathtt{t}$. So if a frame is fully extensional then, for any $\mathtt{t}$ and any v, $\llbracket \mathtt{t}\rrbracket^v = \llbracket \lambda\mathtt{x}(\mathtt{t}\cdot\mathtt{x})\rrbracket^v$ for $\mathtt{x}$ not free in $\mathtt{t}$. This implies, by reasoning similar to Corollary 3.15, that $\bullet = \circ$.

Definition 4.18. Let $\boldsymbol{\eta} = \{\langle \lambda\mathrm{x}.(\mathrm{y}{\cdot}\mathrm{x}), \mathrm{y}\rangle, \langle \mathrm{y}, \lambda\mathrm{x}.(\mathrm{y}{\cdot}\mathrm{x})\rangle\}$ and let $\boldsymbol{\beta\eta} = \boldsymbol{\beta} \cup \boldsymbol{\eta}$.

Theorem 4.19. *If* $\mathtt{t}, \mathtt{s} \in \mathcal{L}_\lambda$ *then,* $\mathtt{t} \Longrightarrow_{\boldsymbol{\beta\eta}} \mathtt{s}$ *iff* $[\![\mathtt{t}]\!]^v \subseteq [\![\mathtt{s}]\!]^v$ *for all fully extensional λ-complete models $\langle F, v\rangle$.*

Proof. It is straightforward to verify (see Remark 4.17) that $\mathtt{t} \Longrightarrow_{\boldsymbol{\beta\eta}} \mathtt{s}$ implies that $[\![\mathtt{t}]\!]^v \subseteq [\![\mathtt{s}]\!]^v$ in all fully extensional, λ-complete models. Conversely, if $\mathtt{t} \not\Longrightarrow_{\boldsymbol{\beta\eta}} \mathtt{s}$ then $[\![\mathtt{t}]\!]^v \nsubseteq [\![\mathtt{s}]\!]^v$ in the canonical model for $\boldsymbol{\beta\eta}$, it is not hard to show that it is λ-complete and fully extensional. □

Definition 4.20. Say that a frame is **combinatorially complete** when for any $h_1, h_2, h_3 \in H$

1. $h_1 = ([\![\mathbf{K}]\!]^v \bullet h_1) \bullet h_2$
2. $(h_1 \bullet h_3) \bullet (h_2 \bullet h_3) = (([\![\mathbf{S}]\!]^v \bullet h_1) \bullet h_2) \bullet h_3$

This is the familiar notion of combinatory completeness as used in characterisations of lambda models in terms of combinatory algebras (e.g. see [12, p.228]). We now get the following result.

Theorem 4.21. *A fully extensional frame (model) is λ-complete if it is combinatorially complete.*

Proof. Again, given Theorem 3.12 we argue partially syntactically. First note that the first two conditions of Definition 4.10 are met if F is combinatorially complete.

Now we argue that if a frame F is combinatorially complete, then for any v,

$$[\![(\mathrm{x}{\cdot}\mathrm{z}){\cdot}\mathrm{y}]\!]^v = [\![((\mathbf{C}{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}]\!]^v$$

Given soundness (Theorem 2.25) and known facts about combinators (e.g. [12, p.25]), if F is combinatorially complete then $[\![(\mathrm{x}{\cdot}\mathrm{z}){\cdot}\mathrm{y}]\!]^v = [\![((\mathbf{C}'{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}]\!]^v$ for some particular complex expression $\mathbf{C}'$ given in terms of $\mathbf{S}$ and $\mathbf{K}$.[7] Moreover, since F is fully extensional, i.e. $[\![\lambda\mathrm{x}(\mathtt{t}{\cdot}\mathrm{x})]\!]^v = [\![\mathtt{t}]\!]^v$ for any $\mathtt{t} \in \mathcal{L}_\lambda$, then

$$[\![\mathbf{C}']\!]^v = [\![\lambda\mathrm{xyz}.((\mathbf{C}'{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}]\!]^v = [\![\lambda\mathrm{xyz}.((\mathrm{x}{\cdot}\mathrm{r}){\cdot}\mathrm{y})]\!]^v = [\![\mathbf{C}]\!]^v$$

and so $[\![(\mathrm{x}{\cdot}\mathrm{z}){\cdot}\mathrm{y}]\!]^v = [\![((\mathbf{C}'{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}]\!]^v = [\![((\mathbf{C}{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}]\!]^v$.

So if F is combinatorially complete then:

$$\begin{array}{rcll} [\![\lambda\mathrm{z}.((\mathrm{x}{\cdot}\mathrm{z}){\cdot}\mathrm{y})]\!]^v & \subseteq & [\![\lambda\mathrm{z}.\big(((\mathbf{C}{\cdot}\mathrm{x}){\cdot}\mathrm{y}){\cdot}\mathrm{z}\big)]\!]^v & \text{by the above} \\ & \subseteq & [\![(\mathbf{C}{\cdot}\mathrm{x}){\cdot}\mathrm{y}]\!]^v & \text{as } F \text{ is fully extensional} \end{array}$$

and so the third condition of 4.10 is met. □

[7] $\mathbf{C}'$ is $\big(\mathbf{S}{\cdot}\big((\mathbf{B}{\cdot}\mathbf{B}){\cdot}\mathbf{S})\big){\cdot}(\mathbf{K}{\cdot}\mathbf{K})\big)$ where $\mathbf{B}$ is short for $(\mathbf{S}{\cdot}(\mathbf{K}{\cdot}\mathbf{S})){\cdot}\mathbf{K}$.

5 Faithfulness

The completeness result 4.15 relates λ theories to *faithful* λ-complete frames. Faithfulness was defined in 2.17 partially syntactically: a faithful frame is one that has a denotation $h \in H$ for every λ-term $\mathtt{t}$.

It is natural to seek a characterisation of faithfulness that does not require explicit reference to the syntax, i.e. a purely semantic one. Can we provide a description, only in terms of H and R, of structural properties a frame must have in order that there is an $h \in H$ to be the denotation of each λ-term? The difficulty lies in the denotation of λ-terms of the form $\lambda\mathtt{x}.\mathtt{s}$. We might know what must hold of H for it to include the denotation of $\mathtt{s}$, but what of $\lambda\mathtt{x}.\mathtt{s}$? $\lambda\mathtt{x}$ acts like a kind of quantifier which binds in $\mathtt{s}$. So the denotation of $\lambda\mathtt{x}.\mathtt{s}$ depends not just on $\mathtt{s}$ but on the denotations of $\mathtt{s}$ for all possible interpretations of $\mathtt{x}$ (assuming it is free in $\mathtt{s}$).

What we have just described is an instance of the more general problem of syntax-free interpretations of quantification and binding (and substitution). We can solve the problem here for the the special case of a syntactic theory of λ-equality (i.e. a $\boldsymbol{\beta}$-sensitive, or λ-complete, theory):[8]

Theorem 5.1. *If an intensional frame F is λ-complete and also for any $S_1, S_2 \subseteq \mathcal{P}(W)$:*

1. $[\![\mathbf{K}]\!]^v, [\![\mathbf{S}]\!]^v, [\![\mathbf{C}]\!]^v \in H$ *(for some/any v),*
2. *if $S_1, S_2 \in H$ then $S_1 \bullet S_2 \in H$.*
3. *if $S_1 \in H$ then $\bigcap_{h\in H}\{w \mid S \circ h \subseteq S_1 \circ h\} \in H$ (i.e. $[\![\lambda\mathtt{y}(\mathtt{x} * \mathtt{y})]\!]^{v[\mathtt{x}\mapsto S_1]} \in H$),*

then F is faithful.

Proof. First notice that condition (1) is independent of v as $\mathbf{K}$, $\mathbf{S}$ and $\mathbf{C}$ are all closed terms. Also note that by Definition 2.17, a frame is faithful when it guarantees an interpretation in H for every term of $\mathcal{L}_\lambda$ (i.e. terms not containing $*$). Condition (2) states that H is closed under $\bullet$. Condition (3) says that if $[\![\mathtt{x}]\!]^v \in H$ then so is $[\![\lambda\mathtt{y}(\mathtt{x} * \mathtt{y})]\!]^v$. Given closure under $\bullet$ this condition (3) could be replaced by the condition that $[\![\lambda\mathtt{xy}(\mathtt{x} * \mathtt{y})]\!]^v \in H$.

We must argue that for any valuation v, $[\![\mathtt{t}]\!]^v \in H$ for all $\mathtt{t} \in \mathcal{L}_\lambda$. We do so by induction on $\mathtt{t}$.

– $\mathtt{t}$ is a variable $\mathtt{x}$. Then $[\![\mathtt{t}]\!]^v \in H$ by the definition of valuations 2.12.

– $\mathtt{t}$ is $\mathtt{s}{\cdot}\mathtt{r}$.Then the result follows by condition (2) and the induction hypothesis.

[8] The mathematical designs of this paper, combined with those of [5, 7], give rise to a more general solution for λ-reduction in [8].

– $\mathtt{t}$ is $\lambda\mathtt{x}.\mathtt{s}$. Then by the induction hypothesis $[\![\mathtt{r}]\!]^v \in H$ for every subterm $\mathtt{r}$ of $\mathtt{s}$. Now $[\mathtt{x}]\mathtt{s}$ contains no free occurrences of $\mathtt{x}$ and is a concatenation, by the $\cdot$ symbol, only of instances of $\mathbf{K}$, $\mathbf{S}$, $\mathbf{C}$ and subterms of $\mathtt{s}$. So by conditions (1), (2) $[\![[\mathtt{x}]\mathtt{s}]\!]^v \in H$. But then by condition (3) and Lemma 2.16.2, $[\![\lambda\mathtt{x}.([\mathtt{x}]\mathtt{s} * \mathtt{x})]\!]^v \in H$. Finally, given λ-completeness we may conclude from Theorems 4.9 and 4.12.1 that $[\![[\mathtt{x}]\mathtt{s} * \mathtt{x}]\!]^v = [\![\mathtt{s}]\!]^v$ and so $[\![\lambda\mathtt{x}.\mathtt{s}]\!]^v \in H$

□

Corollary 5.2. *If an extensional frame F is λ-complete and the three conditions of 5.1 hold, then F is faithful.*

Remark 5.3. It is easy to verify that the converses of 5.1 and 5.2 hold. Moreover, for the case of a fully extensional frame, we know from 4.21 that combinator completeness implies λ-completeness. It is then not hard to see that the conditions of 5.1 become instances of the definition of a syntax-free model of the λ-calculus. For example [12, p.237], condition (3) corresponds to the so-called (although not by [12]) Meyer-Scott axiom.

6 Further work

The methods used here resemble those behind the models of λ-calculus constructed by Engeler, Meyer, Plotkin and Scott (e.g. in [4, 14] and in [1, §18-19]) which are the basis of *graph models*.

The frames presented here have the components W, $\bullet$ and H. Both $\bullet$ and W have an analogue in graph models, and the differences between them and their analogues are not of great significance: it is not hard to associate each graph model with an equivalent extensional frame (see Remark 2.15). The analogue of H in graph models is that denotations are drawn from the powerset of the domain (the analogue of W). The fact that in the models and frames of this paper H can be something other than $\mathcal{P}(W)$ is significant. The completeness theorem 3.12 shows this, for it implies that every consistent λ-theory can be associated with a frame, and yet as shown by Salibra [17] there are λ-theories for which there are no graph models.

The models of this paper separate *expansion* and *reduction* for both $\beta-$ and $\eta-$ as distinct semantic properties of a model. Interestingly, β-reduction and η-expansion are natural features of the models (there is independent evidence that this is natural [13]). η-expansion arises from λ-abstraction and application being defined over the same underlying function $\bullet$. If we use two underlying functions $\bullet$ and $\circ$ instead, where $\bullet \subseteq \circ$, so that λ abstracts over $\circ$ and application applies $\bullet$, then we obtain models free from η-expansion.

We could also reverse this 'trick' so that λ abstracts over the $\bullet$ and application applies $\circ$, and thus obtain models free from β-reduction. This may sound perverse but recall that meta-programming languages — languages that can suspend their own evaluation and/or quote their own syntax — are devoted to switching off β-reduction in a controlled manner, and the connections to modal logic have already been noted, where possible worlds correspond to deeper or shallower levels of suspension or quoting (see MetaML [15] and CMTT [16, 9]). This suggests the possibility of models for a variety of interacting λ-operators over a hierarchy of underlying application relations.

Finally, further work is needed in improving the semantic characterisations, in terms of $\bullet$ and $\circ$, of frames that are Γ-sensitive for interesting Γ. For example, can we provide a helpful semantic characterisation of theories that contain the schema of η-reduction? We can do it in terms of H by specifying that, for any $S \in H$, $\bigcap_{h \in H}\{w \mid w \circ h \subseteq S \bullet h\} \subseteq S$, but it would also be interesting to look for conditions on $\bullet$ and $\circ$ alone that correspond to this, independently of H.

References

[1] Henk P. Barendregt. *The Lambda Calculus: its Syntax and Semantics (revised ed.).* North-Holland, 1984.

[2] Christoph Benzmüller, Chad E. Brown, and Michael Kohlhase. Higher-order semantics and extensionality. *Journal of Symbolic Logic*, 69:1027–1088, 2004.

[3] H.B. Enderton. *A Mathematical Introduction to Logic.* Academic Press, 1972.

[4] Erwin Engeler. Algebras and combinators. *Algebra Universalis*, 13:389–392, 1981.

[5] Michael Gabbay. A proof-theoretic treatment of λ-reduction with cut-elimination: λ-calculus as a logic programming language. *Journal of Symbolic Logic*, 76(2):673–699, 2011.

[6] Michael Gabbay and Murdoch James Gabbay. A simple class of kripke-style models in which logic and computation have equal standing. In Edmund M. Clarke and Andrei Voronkov, editors, *LPAR (Dakar)*, volume 6355 of *Lecture Notes in Computer Science*, pages 231–254. Springer, 2010.

[7] Murdoch J. Gabbay. Semantics out of context: nominal absolute denotations for first-order logic and computation. 2012. Submitted; available as arXiv preprint arxiv.org/abs/1305.6291.

[8] Murdoch J. Gabbay and Michael J. Gabbay. Representation and duality of the untyped lambda-calculus in nominal lattice and topological semantics, with a proof of topological completeness. 2012. Submitted; available as arXiv preprint arxiv.org/abs/1305.5968.

[9] Murdoch J. Gabbay and Aleksandar Nanevski. Denotation of contextual modal type theory (CMTT): syntax and metaprogramming. *Journal of Applied Logic*, 11:1–29, March 2013.

[10] Robert Goldblatt. *Logics of Time and Computation.* Number 7 in CSLI Lecture Notes. Center for the Study of Language and Information, 2. edition, 1992.

[11] Leon Henkin. Completeness in the theory of types. *Journal of Symbolic Logic*, 15:81–91, 1950.

[12] J. Roger Hindley and Jonathan P. Seldin. *Lambda-Calculus and Combinators, An Introduction.* Cambridge University Press, 2nd edition, 2008.

[13] C. Barry Jay and Neil Ghani. The virtues of eta-expansion. *Journal of Functional Programming*, 5(2):135–154, April 1995.

[14] Albert R. Meyer. What is a model of the lambda calculus? *Information and Control*, 1(52):87–122, 1982.

[15] Eugenio Moggi, Walid Taha, Zine-El-Abidine Benaissa, and Tim Sheard. An idealized metaml: Simpler, and more expressive. In *ESOP '99: Proc. of the 8th European Symposium on Programming Languages and Systems*, volume 1576 of *Lecture Notes in Computer Science*, pages 193–207. Springer, 1999.

[16] Aleksandar Nanevski, Frank Pfenning, and Brigitte Pientka. Contextual modal type theory. *ACM Transactions on Computational Logic (TOCL)*, 9(3):1–49, 2008.

[17] Antonino Salibra. Topological incompleteness and order incompleteness of the lambda calculus. *ACM Trans. Comput. Logic*, 4(3):379–401, 2003.

[18] Peter Selinger. Order-incompleteness and finite lambda reduction models. *Theoretical Computer Science*, 309(1):43–63, 2003.

[19] Stewart Shapiro. *Foundations without foundationalism: a case for second-order logic.* Number 17 in Oxford logic guides. Oxford University Press, 2000.

Received February 2014

Computer-Aided Discovery and Categorisation of Personality Axioms

Simon Kramer
SK-R&D Ltd liab. Co
simon.kramer@a3.epfl.ch

Abstract

We propose a computer-algebraic, order-theoretic framework based on intuitionistic logic for the computer-aided discovery of personality axioms from personality-test data and their mathematical categorisation into formal personality theories in the spirit of F. Klein's *Erlanger Programm* for geometrical theories. As a result, formal personality theories can be automatically generated, diagrammatically visualised, and mathematically characterised in terms of categories of invariant-preserving transformations in the sense of Klein and category theory. Our personality theories and categories are induced by implicational invariants that are ground instances of intuitionistic implication, which we postulate as axioms. In our mindset, the essence of personality, and thus mental health and illness, is its invariance. The truth of these axioms is algorithmically extracted from histories of partially-ordered, symbolic data of observed behaviour. The personality-test data and the personality theories are related by a Galois-connection in our framework. As data format, we adopt the format of the symbolic values generated by the Szondi-test, a personality test based on L. Szondi's unifying, depth-psychological theory of fate analysis.

Keywords: Applied Order Theory, Computational and Mathematical Depth Psychology, Data Mining, Diagrammatic Reasoning, Fuzzy Implication, Intuitionistic Logic, Logical and Visual Data Analytics, Personality Tests, Szondi.

1 Introduction

In 1872, Felix Klein, full professor of mathematics at the University of Erlangen at age 23, presented his influential *Erlanger Programm* [11, 12] on the classification and characterisation of geometrical theories by means of group theory. That is, Klein put

For the technical-report version of this article, see [13].

forward the thesis that every geometrical theory could be characterised by an associated group of geometrical transformations that would leave invariant the essential properties of the geometrical objects of that theory. These essential properties are captured by the axioms that define the theory. As a result, geometrical theories could be classified in terms of their associated transformation groups. According to [9], Klein's *Erlanger Programm* "is regarded as one of the most influential works in the history of geometry, and more generally mathematics, during the half-century after its publication in 1872."[1]

In this paper and in the spirit of the *Erlanger Programm* for geometrical theories, we propose a computer-algebraic,[2] order-theoretic framework based on intuitionistic logic [18] for the computer-aided discovery of personality axioms from personality-test data and their mathematical categorisation into formal personality theories. Each one of the resulting intuitionistic personality theories is an (order-theoretic) *prime filter* [4] in our framework. As our contribution, formal personality theories can be automatically generated, diagrammatically visualised, and mathematically characterised in terms of categories of invariant-preserving transformations in the sense of Klein and category theory [17]. That is, inspired by and in analogy with Klein, we put forward the thesis that every personality theory can be characterised by an associated category of personality transformations that leave invariant the essential properties of "the personality objects"—the people, represented by their personality-test data—of that theory.

So, the reason for why psychologists and logicians or mathematicians should be interested in this paper is, as for psychologists, our automatic generation and diagrammatic visualisation of formal personality theories and, as for logicians and mathematicians, our mathematical characterisation of such logical theories. Of course, our ideal readership is an interdisciplinary community of logically and mathematically inclined psychologists as well as psychologically inclined logicians and mathematicians, since our methodology is mathematical (formal logical) and our application domain psychology (personality assessment). Consider that arguably, psychology still largely is a pre-scientific discipline: besides statistics, psychological methodology does not employ much of mathematics. (And statistics often is not even part of mathematics departments.) Certainly, psychology is not an exact science — not yet. Here and now, our formal framework is meant as a contribution towards practicing psychological research with the methods of the exact sciences, for obvious ethical reasons, and as a contribution towards the mathematical systematisation of the academic discipline of psychology, particularly in the area of test-based person-

[1] We add that in physics, the existence of certain transformation groups for mechanics and electromagnetism led Albert Einstein to discover his theory of special relativity.

[2] in the sense of *symbolic* as opposed to numeric computation

ality theories. Interesting psychological insights resulting from this formal practice are the discovered personality axioms (as exemplified by Table 2 and 3; consider also the accompanying explanations). The main logical insight is that these axioms turn out to be intuitionistic implications, which by their very nature express invariants. (In science, we are interested in laws, things that are eternally true, that is, invariant. The goal of any scientific quest is to discover these laws.)

An important difference in our psychological context of personality theories to Klein's geometrical context is that actually no formal personality theory in the strict axiomatic sense exists, whereas Klein could characterise a variety of *existing,* axiomatic theories of geometry. Before being able to categorise personality theories, we thus must first formally *define* them. As said, we shall do so by discovering their defining axioms from personality-test data with the aid of computers. Our personality theories and categories are then automatically induced by implicational invariants that hold throughout that test data and that are ground instances of intuitionistic implication, which we postulate as axioms. In our mindset, the essence of personality, and thus mental health and illness, is its invariance. So for every person, represented by her personality-test result P—the data—we automatically generate her associated

1. personality *theory* $\{P\}^{\triangleleft}$ of simple implicational invariants and
2. personality *category* $\mathbf{T}_{\{P\}^{\triangleleft}}$ of theory-preserving transformations.

(We are actually able to carry out this construction for whole *sets* of personality-test results, either of different people or of one and the same person.) More precisely, the truth of these axioms is algorithmically extracted from histories of partially-ordered, symbolic data of the person's observed test behaviour. As an example, Table 3 visualises as black cells all axioms that were algorithmically extracted from Table 2 and that can be added to intuitionistic propositional logic to constitute such an axiomatically defined theory for the person in question. These axioms are then also enumerated as formulas and their psychological meaning explained. Of course, any set of people with the same extracted axioms forms a natural cluster of people with the same personality theory. Note that our axioms have an implicational form in order to conform with the standard of Hilbert-style axiomatisations [7], which in our order-theoretic framework can be cast as a simple closure operator. Another difference in our context is that contrary to Klein, who worked with transformation groups, we work with more general transformation *monoids,* and thus transformation categories. The reason is that contrary to Klein's geometrical context, in which transformations are invertible, transformations in the psychological context need not be invertible.

In our order-theoretic framework, personality-test data and personality theories are related by a Galois-connection $({}^{\triangleright}, {}^{\triangleleft})$ [4, Chapter 7]. As data format, we adopt—without loss of generality—the format of the symbolic values, called *Szondi personality profiles (SPPs),* generated by the Szondi-test [22], a personality test based on L. Szondi's unifying, depth-psychological theory of fate analysis [23]. An SPP can be conceived as a tuple of eight, so-called *signed factors* whose signatures can in turn take twelve values. We stress that our framework is independent of any personality test. It simply operates on the result values that such tests generate. Our choice of the result values of the Szondi-test is motivated by the fact that SPPs just happen to have a finer structure than other personality-test values that we are aware of, and so are perhaps best suited to play the illustrative role for which we have chosen them here. (See also [14].)

The remaining part of this paper is structured as follows: in Section 2, we present the part of our framework for the computer-aided discovery of personality axioms from personality-test data, and in Section 3, the part for their mathematical categorisation into formal personality theories.

2 Axiom discovery

In this section, we present the part of our framework for the computer-aided discovery of personality axioms from personality-test data. This is the data-mining and the logical and visual data-analytics part of our contribution.

We start with defining the format of the data on which we perform our data-mining and data-analytics operations. As announced, it is the format of the symbolic values, called *Szondi personality profiles (SPPs),* generated by the Szondi-test [22]. We operate on finite sequences thereof. In diagnostic practice, these test-result sequences are usually composed of 10 SPPs [15].

Definition 1 (The Szondi-Test Result Space)**.** Let us consider the Hasse-diagram [4] in Figure 1 of the partially ordered set of *Szondi's twelve signatures* [22] of human reactions, which are:

- approval: from strong $+!!!$, $+!!$, and $+!$ to weak $+$;
- indifference/neutrality: 0;
- rejection: from weak $-$, $-!$, and $-!!$ to strong $-!!!$; and
- ambivalence: $\pm^{!}$ (approval bias), $\pm$ (no bias), and $\pm_{!}$ (rejection bias).

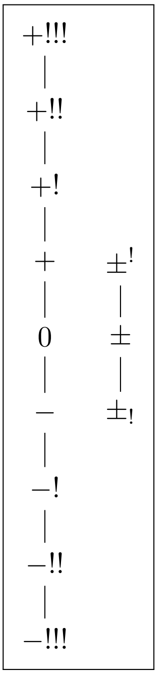


Figure 1: Hasse-diagram of Szondi's signatures

(Szondi calls the exclamation marks in his signatures *quanta.*)

Further let us call this set of signatures $\mathbb{S}$, that is,

$$\mathbb{S} := \{\, -!!!, -!!, -!, -, 0, +, +!, +!!, +!!!, \pm_{!}, \pm, \pm^{!} \,\}.$$

Now let us consider *Szondi's eight factors and four vectors* of human personality [22] as summarised in Table 1. (Their names are of clinical origin and need not concern us here.) And let us call the set of factors $\mathbb{F}$, that is,

$$\mathbb{F} := \{\, \mathsf{h}, \mathsf{s}, \mathsf{e}, \mathsf{hy}, \mathsf{k}, \mathsf{p}, \mathsf{d}, \mathsf{m} \,\}.$$

Then,

- $\mathrm{SPP} := \{\ ((\mathsf{h}, s_1), (\mathsf{s}, s_2), (\mathsf{e}, s_3), (\mathsf{hy}, s_4), (\mathsf{k}, s_5), (\mathsf{p}, s_6), (\mathsf{d}, s_7), (\mathsf{m}, s_8)) \mid s_1, \ldots, s_8 \in \mathbb{S}\ \}$

  is the set of Szondi's personality profiles;

- $\langle\, \mathrm{SPP}^{+}, \star \,\rangle$ is the free semigroup on the set SPP^{+} of all finite sequences of

Vector	**Factor**	**Signature**	
		+	−
S (Id)	h (love)	physical love	platonic love
	s (attitude)	(proactive) activity	(receptive) passivity
P (Super-Ego)	e (ethics)	ethical behaviour	unethical behaviour
	hy (morality)	immoral behaviour	moral behaviour
Sch (Ego)	k (having)	having more	having less
	p (being)	being more	being less
C (Id)	d (relations)	unfaithfulness	faithfulness
	m (bindings)	dependence	independence

Table 1: Szondi's factors and vectors

SPPs with $\star$ the (associative) concatenation operation on SPP^+ and

$$\begin{aligned}\mathrm{SPP}^+ &:= \bigcup_{n\in\mathbb{N}\setminus\{0\}} \mathrm{SPP}^n\\ \mathrm{SPP}^{1+n} &:= \mathrm{SPP}^1 \times \mathrm{SPP}^n\\ \mathrm{SPP}^1 &:= \mathrm{SPP}\,;\end{aligned}$$

- $\mathcal{STR} := \langle\,\mathrm{SPP}^+, \sqsubseteq\rangle$ is our *Szondi-Test Result Space,* where the suffix partial order $\sqsubseteq$ on SPP^+ is defined such that for every $P, P' \in \mathrm{SPP}^+$, $P \sqsubseteq P'$ if and only if $P = P'$ or there is $P'' \in \mathrm{SPP}^+$ such that $P = P'' \star P'$.

As an example of an SPP, consider the *norm profile* for the Szondi-test [22]:

$$((\mathsf{h},+),(\mathsf{s},+),(\mathsf{e},-),(\mathsf{hy},-),(\mathsf{k},-),(\mathsf{p},-),(\mathsf{d},+),(\mathsf{m},+))$$

Spelled out, the norm profile describes the personality of a human being who approves of physical love, has a proactive attitude, has unethical but moral behaviour, wants to have and be less, and is unfaithful and dependent.

Those SPP-sequences that have been generated by a Szondi-test(ee) are our histories of partially-ordered, symbolic data of observed behaviour that we announced in the introduction. Table 2 displays an example of such an SPP-sequence: it is the so-called foreground profile of a 49-year old, male physician and psycho-hygienist and is composed of 10 subsequent SPPs [21, Page 182–184].

Fact 1 (Prefix closure of $\sqsubseteq$)**.** For every $P, P', P'' \in \mathrm{SPP}^+$,

$$P \sqsubseteq P' \text{ implies } P'' \star P \sqsubseteq P'$$

Nr.	S		P		Sch		C	
	h	s	e	hy	k	p	d	m
1	−	0	±	±	±	±	0	+
2	−	0	+	±	±	+	0	+
3	−	−	±	±	±	+	+	±
4	−	−	±	+	+	+	0	+
5	−	0	0	+	±	±	0	+
6	−	0	±	±	±	±	+	±
7	−	0	±	±	±	+	0	+
8	−	−	0	±	+	+	+	±
9	−	0	±	±	±	±	0	+
10	−	0	0	±	±	+	0	+

Table 2: A Szondi-test result (say P)

Proof. By inspection of definitions. □

We continue to define the closure operator by which we generate our intuitionistic personality theories from personality-test data in the previously-defined format. Our personality theories are intuitionistic, because such theories can be interpreted over partially-ordered state spaces—such as our $\mathcal{STR}$—such that a sentence is true in the current state by definition if and only if the sentence is true in all states that are accessible from the current state by means of the partial order [16, 18]. In other words, the truth of such sentences is forward-invariant, which is precisely the property of sentences that we are looking for.

Definition 2 (A closure operator for intuitionistic theories)**.** Let

$$\mathbb{A} := \{\, \mathsf{h}s_1, \mathsf{s}s_2, \mathsf{e}s_3, \mathsf{hy}s_4, \mathsf{k}s_5, \mathsf{p}s_6, \mathsf{d}s_7, \mathsf{m}s_8 \mid s_1, \ldots, s_8 \in \mathbb{S} \,\}$$

be our set of atomic statements, and

$$\mathcal{L}(\mathbb{A}) \ni \phi ::= A \mid \phi \land \phi \mid \phi \lor \phi \mid \neg\phi \mid \phi \rightarrow \phi \quad \text{for } A \in \mathbb{A}$$

our logical language over $\mathbb{A}$, that is, the set of statements ϕ constructed from the atomic statements A and the intuitionistic logical connectives $\land$ (conjunction, pronounced “and”), $\lor$ (disjunction, pronounced “or”), $\neg$ (negation, pronounced “henceforth not”), and $\rightarrow$ (implication, pronounced “whenever—then”). As usual, we can macro-define falsehood as $\bot := A \land \neg A$ and truth as $\top := \neg\bot$.

Further let

$\Gamma_0 := \{$

- $\phi \to (\phi' \to \phi)$
- $(\phi \to (\phi' \to \phi'')) \to ((\phi \to \phi') \to (\phi \to \phi''))$
- $(\phi \wedge \phi') \to \phi$
- $(\phi \wedge \phi') \to \phi'$
- $\phi \to (\phi' \to (\phi \wedge \phi'))$
- $\phi \to (\phi \vee \phi')$
- $\phi' \to (\phi \vee \phi')$
- $(\phi \to \phi') \to ((\phi'' \to \phi') \to ((\phi \vee \phi'') \to \phi'))$
- $\bot \to \phi\ \}$

be our (standard) set of intuitionistic *axiom schemas.*

Then, $\mathrm{Cl}(\emptyset) := \bigcup_{n\in\mathbb{N}} \mathrm{Cl}^n(\emptyset)$, where for every $\Gamma \subseteq \mathcal{L}(\mathbb{A})$:

$$\begin{array}{rcl} \mathrm{Cl}^0(\Gamma) & := & \Gamma_0 \cup \Gamma \\ \mathrm{Cl}^{n+1}(\Gamma) & := & \mathrm{Cl}^n(\Gamma) \cup \\ & & \{\, \phi' \mid \{\phi, \phi \to \phi'\} \subseteq \mathrm{Cl}^n(\Gamma)\,\} \quad (\textit{modus ponens}, \text{MP}) \end{array}$$

We call $\mathrm{Cl}(\emptyset)$ our *base theory,* and $\mathrm{Cl}(\Gamma)$ a Γ-*theory* for any $\Gamma \subseteq \mathcal{L}(\mathbb{A})$.

The following standard fact asserts that we have indeed defined a closure operator. We merely state it as a reminder, because we shall use it in later proof developments. The term $2^{\Gamma}_{\text{finite}}$ denotes the set of all finite subsets of the set Γ.

Fact 2. The mapping $\mathrm{Cl} : 2^{\mathcal{L}(\mathbb{A})} \to 2^{\mathcal{L}(\mathbb{A})}$ is a *standard consequence operator,* that is, a *substitution-invariant compact closure operator:*

1. $\Gamma \subseteq \mathrm{Cl}(\Gamma)$ (extensivity)
2. if $\Gamma \subseteq \Gamma'$ then $\mathrm{Cl}(\Gamma) \subseteq \mathrm{Cl}(\Gamma')$ (monotonicity)
3. $\mathrm{Cl}(\mathrm{Cl}(\Gamma)) \subseteq \mathrm{Cl}(\Gamma)$ (idempotency)
4. $\mathrm{Cl}(\Gamma) = \bigcup_{\Gamma' \in 2^{\Gamma}_{\text{finite}}} \mathrm{Cl}(\Gamma')$ (compactness)
5. $\sigma[\mathrm{Cl}(\Gamma)] \subseteq \mathrm{Cl}(\sigma[\Gamma])$ (substitution invariance),

where σ designates an arbitrary propositional $\mathcal{L}(\mathbb{A})$-substitution.

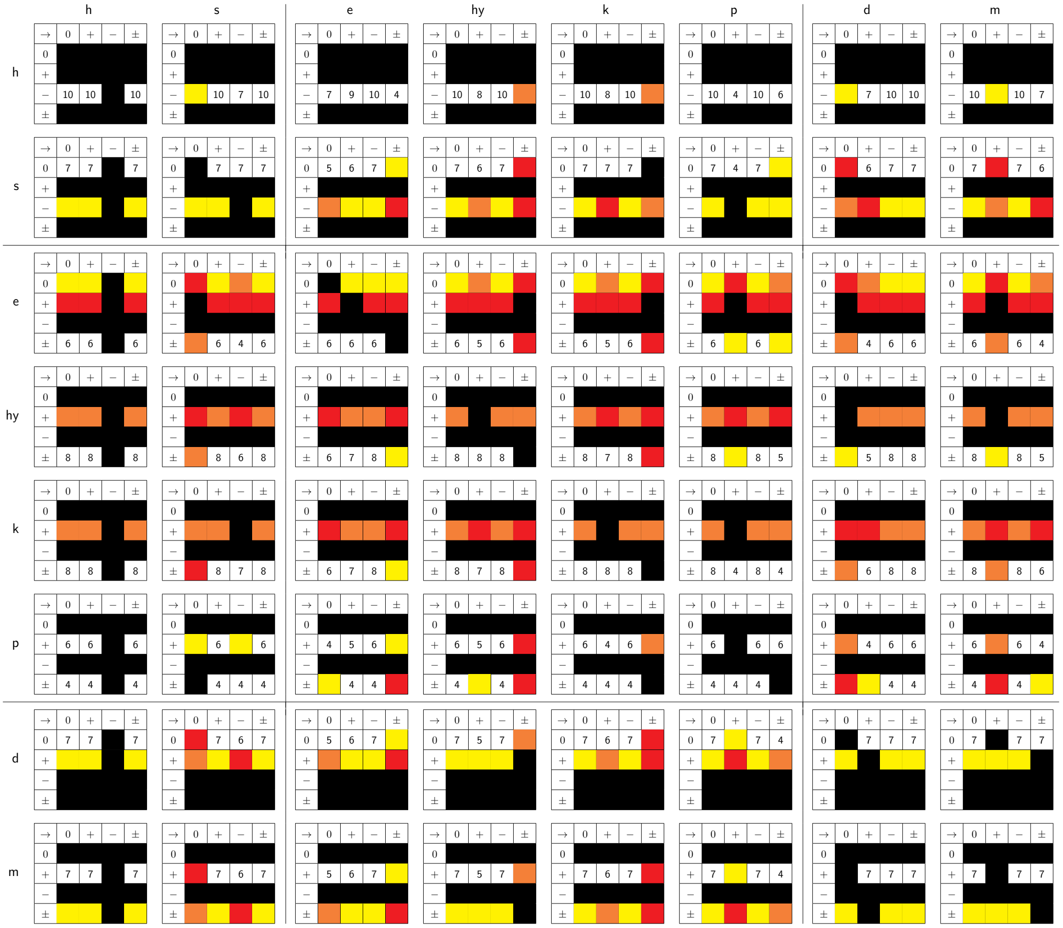


Table 3: The diagram of $\mathcal{I}(P)$ as extracted from the P in Table 2

Proof. For (1) to (4), inspect the inductive definition of Cl. And (5) follows from our definitional use of axiom *schemas*.[3] □

Note that in the sequel, ":iff" abbreviates "by definition, if and only if," and

$$\begin{aligned} \Phi \vdash_\Gamma \phi \quad &:\text{iff} \quad \Phi \subseteq \mathrm{Cl}(\Gamma) \text{ implies } \phi \in \mathrm{Cl}(\Gamma) \\ \vdash_\Gamma \phi \quad &:\text{iff} \quad \emptyset \vdash_\Gamma \phi\,. \end{aligned}$$

[3] Alternatively to axiom schemas, we could have used axioms together with an additional substitution-rule set $\{\sigma[\phi] \mid \phi \in \mathrm{Cl}^n(\Gamma)\}$ in the definiens of $\mathrm{Cl}^{n+1}(\Gamma)$.

We continue to define what we mean by our simple implicational invariants announced in the introduction. As announced there, these invariants are ground instances of intuitionistic implication, by which we mean that they are of the visually tractable, diagrammatic form $A \rightarrow A'$ rather than being of the more general, not generally visually tractable form $\phi \rightarrow \phi'$. As an example of what we mean by visually tractable, diagrammatic form, consider Table 3. For a given SPP-sequence P, we postulate the algorithmically extracted set $\mathcal{I}(P)$ of these invariants that hold throughout P (see Definition 3) as the axioms of the personality theory $\mathrm{Cl}(\mathcal{I}(P))$ that we associate with P. These axioms thus capture those *logical dependencies* between signed factors that are invariant in P in the sense of holding throughout P. The algorithm for this axiom extraction and visualisation is displayed in Listing 1 and will be explained shortly. Note that given that these invariants hold throughout a sequence that has been generated by an iterated procedure, that is, an iterated execution of the Szondi-test, they can also be understood as *loop invariants,* which is a core concept in the science of computer programming [6]. So our algorithm for finding psychological invariants can actually also be understood and even be used as a method for inferring loop invariants from program execution traces in computer science.

Definition 3 (Simple implicational invariants)**.** Let the mapping $\mathrm{p} : \mathrm{SPP} \rightarrow \mathcal{L}(\mathbb{A})$ be such that

$$\begin{aligned}\mathrm{p}(((\mathsf{h}, s_1), (\mathsf{s}, s_2), (\mathsf{e}, s_3), (\mathsf{hy}, s_4), (\mathsf{k}, s_5), (\mathsf{p}, s_6), (\mathsf{d}, s_7), (\mathsf{m}, s_8))) = \\ \mathsf{h}s_1 \wedge \mathsf{s}s_2 \wedge \mathsf{e}s_3 \wedge \mathsf{hy}s_4 \wedge \mathsf{k}s_5 \wedge \mathsf{p}s_6 \wedge \mathsf{d}s_7 \wedge \mathsf{m}s_8 \,.\end{aligned}$$

Then, define the mapping $\mathcal{I} : \mathrm{SPP}^+ \rightarrow 2^{\mathcal{L}(\mathbb{A})}$ of *simple implicational invariants* such that for every $P \in \mathrm{SPP}^+$,

$$\mathcal{I}(P) := \{\, A \rightarrow A' \mid \begin{array}{c}\text{for every } P' \in \mathrm{SPP}^+\text{, if } P \sqsubseteq P' \\ \text{then } \mathrm{p}(\pi_1(P')) \vdash_\emptyset A \rightarrow A'\end{array} \,\}\,,$$

where $\pi_1 : \mathrm{SPP}^+ \rightarrow \mathrm{SPP}$ is projection onto the first SPP component.

Notice the three implications "if—then," $\vdash$, and $\rightarrow$ of different logical level, and note that we use "if—then" and "implies" synonymously. This definition can be cast into an algorithm of linear complexity in the length of P, for example as described by the Java-program displayed in Listing 1, and the result $\mathcal{I}(P)$ of its computation diagrammatically displayed as in Table 3. The on-line Szondi-test [15] also uses this program as a subroutine. Lines starting with "`//` " are comments. Notice that every loop in the program has fixed complexity, and that we simply process the head of P—the first profile in P—in Line 2–30 and then recur on the tail of the remaining

```
public void update (Vector<Signature[]> profiles) {
  // 1. CALCULATION OF MATERIAL IMPLICATIONS
  // in the first profile
  Signature[] fstp = profiles.firstElement();
  // consequent-oriented processing (consequent loop),
  // round-robin treatment of each factor as consequent
  for (int c=0; c<8; c++) {
    Signature cmodq = moduloQuanta(fstp[c]);
    // 1.1 EVERYTHING IMPLIES TRUTH (antecedent loop),
    // round-robin treatment of each factor as antecedent
    for (int a=0; a<8; a++) {
      // signature-value loop
      for (int v=0; v<4; v++) {
        // discount corresponding table-cell value
        (factors[a][c]).signatures[v][code(cmodq)]--;
      }
    }
    // 1.2 FALSEHOOD IMPLIES EVERYTHING—ALSO FALSEHOOD;
    // everything is: all other consequent signatures
    for (Signature cc : coSet(cmodq)) {
      // round-robin treatment of each factor as antecedent
      for (int a=0; a<8; a++) {
        // false is: all other antecedent signatures
        for (Signature ca : coSet(moduloQuanta(fstp[a]))) {
          // discount corresponding table-cell value
          (factors[a][c]).signatures[code(ca)][code(cc)]--;
        }
      }
    }
  }
  // 2. CALCULATION OF INTUITIONISTIC IMPLICATIONS:
  // forward invariance of material implications
  if (profiles.size()>1) {
    // garbage-collect the processed profile
    profiles.remove(0);
    // recursively descend on the remaining profiles
    update(profiles);
  }
}
```

Listing 1: Update algorithm

	A	A'	$A \supset A'$
1.	false	false	true
2.	false	true	true
3.	true	false	false
4.	true	true	true

Table 4: Material implication $\supset$

profiles in P in Line 31–38. The loop-nesting depth is four. The program updates a table—called *factors* in Listing 1—of eight times eight subtables—called *signatures* in Listing 1—of four times four content cells as displayed in Table 3, each of whose cells is initialised with a value equal to the length of P (e.g., 10). To update this table means to discount the initial value of its cells according to the following strategy inspired by Kripke's model-theoretic interpretation of intuitionistic implication as forward invariance of material implication [16, 18] and adapted to our setting in Definition 3:

1. Calculate all material implications in the first profile in P, called *profiles* in Listing 1, according to the definition of material implication recalled in Table 4. There, Line 1, 2, and 4 can be summarised by the slogan "Everything implies truth" and Line 1 and 2 by the slogan "Falsehood implies everything." In Listing 1, these well-known slogans correspond to the meaning of our code in Line 9–17 and Line 18–29, respectively. There, the function *moduloQuanta* simply returns its argument signature without quanta for graphical tractability, the function *code* the subtable line number of its argument, and the function *coSet* the set of all plain signatures (those signatures without quanta) minus the argument signature. For example, applying

   - *moduloQuanta* to the signature +!!! returns the signature +,
   - *code* to the signature + returns the line number 1, and
   - *coSet* to the signature + returns the set of signatures $\{-, 0, \pm\}$.

2. Then calculate those material implications that are actually even intuitionistic implications by recurring on the tail of P.

On termination of program execution, each table cell that corresponds to an intuitionistic implication will contain the number 0 and be painted black. Cells containing the number 1 will be painted red (and correspond to intuitionistic implications of the tail of P), those containing the number 2 will be painted orange, and those containing the number 3 yellow. Table cells containing other numbers will not be painted for lack of relevance and thus will just display the number of missing discounts as distance to count as intuitionistic implications.

Observe in Table 3 that the whole diagonal from the top left corner down to the bottom right corner is painted black. This state of affairs reflects the reflexivity property $\vdash_\Gamma \phi \rightarrow \phi$ of intuitionistic implication. Similarly, if both a cell representing some formula $A \rightarrow A'$ as well as another cell representing some formula $A' \rightarrow A''$ are painted black then the cell representing the formula $A \rightarrow A''$ will also be painted black. Consider the following four example triples:

- $\mathsf{e}+ \rightarrow \mathsf{s}0$, $\mathsf{s}0 \rightarrow \mathsf{k}\pm$, and $\mathsf{e}+ \rightarrow \mathsf{k}\pm$;
- $\mathsf{k}+ \rightarrow \mathsf{s}-$, $\mathsf{s}- \rightarrow \mathsf{p}+$, and $\mathsf{k}+ \rightarrow \mathsf{p}+$;
- $\mathsf{p}\pm \rightarrow \mathsf{s}0$, $\mathsf{s}0 \rightarrow \mathsf{k}\pm$, and $\mathsf{p}\pm \rightarrow \mathsf{k}\pm$;
- $\mathsf{m}\pm \rightarrow \mathsf{d}+$, $\mathsf{d}+ \rightarrow \mathsf{hy}\pm$, and $\mathsf{m}\pm \rightarrow \mathsf{hy}\pm$.

This state of affairs reflects the transitivity property of intuitionistic implication, which for general formulas ϕ and ϕ' is:

$$\text{if } \vdash_\Gamma \phi \rightarrow \phi' \text{ and } \vdash_\Gamma \phi' \rightarrow \phi'' \text{ then } \vdash_\Gamma \phi \rightarrow \phi''.$$

Of course, reflexivity and transitivity are two logical properties, which will show up in the diagram of any $\mathcal{I}(P)$. (In so far, these properties also reflect an axiomatic redundancy of $\mathcal{I}(P)$, which however is not our concern here.)

In contrast, the following properties displayed in Table 3 are *psychological* in that they are proper to the personality profile displayed in Table 2, from which they have been extracted, namely:

Vacuous implications This class of intuitionistic implications is visually characterised by a horizontal or vertical line of black cells throughout the whole diagram width and diagram height, respectively. In the diagram displayed in Table 3, there is a single vertical line of such implications, which says:

Whenever something is true then $\mathsf{h}-$ is true.

This simply holds because $\mathsf{h}-$ is true throughout the whole test result in Table 2, and thus the logical slogan "Everything implies truth" applies.

The other, that is, the horizontal lines of vacuous intuitionistic implications hold due to the other logical slogan "Falsehood implies everything."

Non-vacuous implications This class of intuitionistic implications is visually characterised by isolated black cells. They are the psychologically truly interesting implications. They are, from the top left to the bottom right of the diagram in Table 3, and together with their rough psychological meaning in Szondi's system (recall Table 1 and, if need be, consult [22]):

1. $\mathsf{s0} \rightarrow \mathsf{k}\pm$. Whenever the testee is inactive (externally) then he has internal compulsive behaviour (e.g., is experiencing a dilemma). See for example the below Item 3 and 6, where these two implicationally related reactions also appear conjunctively.
2. $\mathsf{s}- \rightarrow \mathsf{p}+$. Whenever the testee is receptively passive (e.g., masochism) then he wants to be more than he actually is (e.g., megalomania). See for example the below Item 5, where these two implicationally related reactions also appear conjunctively.
3. $\mathsf{e}+ \rightarrow (\mathsf{s0} \land \mathsf{hy}\pm \land \mathsf{k}\pm \land \mathsf{p}+ \land \mathsf{d0} \land \mathsf{m}+)$: Whenever the testee has ethical behaviour then he
   (a) is inactive ($\mathsf{s0}$). That is, inactivity is a necessary condition for the testee's ethical behaviour, and thus his ethical behaviour is in a behavioural (not logical) sense vacuous.
   (b) is morally ambivalent ($\mathsf{hy}\pm$). Thus the testee's ethical behaviour need not be moral. Indeed, inactivity need not be moral.
   (c) has internal compulsive behaviour ($\mathsf{k}\pm$). Maybe the testee's inactivity is due to an internally experienced dilemma?
   (d) wants to be more than he is ($\mathsf{p}+$). The testee's inactivity may not be conducive to the fulfilment of his desire, but his desire may well be co-determined by his inactivity.
   (e) is faithfully indifferent ($\mathsf{d0}$). Indeed, faithfulness (in a general sense) and ethics may be experienced as orthogonal issues.
   (f) approves of bindings in his relationships ($\mathsf{m}+$). Thus for the testee, bindings but not necessarily their faithfulness are ethical.
4. $\mathsf{hy}+ \rightarrow (\mathsf{d0} \land \mathsf{m}+)$: Whenever the testee has immoral behaviour then he
   (a) is faithfully indifferent ($\mathsf{d0}$).
   (b) approves of bindings in his relationships ($\mathsf{m}+$).

   Thus the testee's faithfulness indifference as well as his binding attitude is stable with respect to ethics and immorality.
5. $\mathsf{k}+ \rightarrow (\mathsf{s}- \land \mathsf{p}+)$: Whenever the testee wants to have more than he has then he

   (a) is receptively passive ($\mathsf{s}-$).
   (b) wants to be more than he is ($\mathsf{p}+$).

   Again, the testee's receptive passivity may not be conducive to the fulfilment of his desires, but his desires may well be co-determined by his receptive passivity.

6. $\mathsf{p}\pm \to (\mathsf{s0} \wedge \mathsf{k}\pm)$: Whenever the testee is ambivalent with respect to being more or less than he is then he
   (a) is inactive ($\mathsf{s0}$). The testee's ambivalence may well be a deeper dilemma that is the cause of his activity blockage.
   (b) has internal compulsive behaviour ($\mathsf{k}\pm$). This could be a confirmation of the testee's suspected dilemma.
7. $\mathsf{d0} \to \mathsf{m}+$. Whenever the testee is faithfully indifferent then he approves of bindings in his relationships. See for example the above Item 3 and 4, where these two implicationally related reactions also appear conjunctively.
8. $\mathsf{d}+ \to \mathsf{hy}\pm$. Whenever the testee is unfaithful then he is morally ambivalent. See for example the below Item 10, where these two implicationally related reactions also appear conjunctively.
9. $\mathsf{m}+ \to \mathsf{d0}$. Whenever the testee approves of bindings in his relationships then he is faithfully indifferent. For example see Item 3 and 4, where these two implicationally related reactions also appear conjunctively.
10. $\mathsf{m}\pm \to (\mathsf{hy}\pm \wedge \mathsf{d}+)$: Whenever the testee is ambivalent in his attitude towards bindings in his relationships then he
    (a) is morally ambivalent ($\mathsf{hy}\pm$).
    (b) is unfaithful ($\mathsf{d}+$).

Observe that from the above invariants in the given P, we can deduce that:

$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{s0} \vee \mathsf{p}\pm) \to \mathsf{k}\pm$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{s}- \vee \mathsf{k}+) \to \mathsf{p}+$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{hy}+) \to (\mathsf{d0} \wedge \mathsf{m}+)$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{hy}+ \vee \mathsf{m}+) \to \mathsf{d0}$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{hy}+ \vee \mathsf{d0}) \to \mathsf{m}+$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{p}\pm) \to \mathsf{s0}$$
$$\vdash_{\mathcal{I}(P)} (\mathsf{e}+ \vee \mathsf{d}+ \vee \mathsf{m}\pm) \to \mathsf{hy}\pm$$

From our diagrammatic reasonings, it becomes clear that the signed factor $\mathsf{e}+$ and to a lesser extent the signed factor $\mathsf{hy}+$ are the two most important *causal factors* in P—and thus for the testee represented by P—in the following sense:

1. these factors are non-vacuously implied by no other signed factor, but

2. they individually and non-vacuously imply most other signed factors.

Thus the testee's personality is determined to a large extent by these two signed factors, in spite of the fact that they only occur in P once and twice, respectively!

Diagrammatic reasoning in a couple Given two (or more) SPP-sequences P and P', we can compute their axiom bases $\mathcal{I}(P)$ and $\mathcal{I}(P')$, and visualise them as diagrams. We can then graphically compute the join and the meet of P and P', that is, the union and the intersection of $\mathcal{I}(P)$ and $\mathcal{I}(P')$, respectively, by simply superposing the two diagrams as printed on overhead-projector foils, and then adding their cells and pinpointing their common cells on a third and fourth superimposed foil, respectively. Of course, this graphical computation can instead also be programmed on a computer (e.g., for studying groups).

Conjunctive implicational invariants As indicated, our algorithmically extracted implicational invariants $A \to A'$ are simple in that they have a single atomic antecedent A and a single atomic consequent A'. An interesting generalisation of these simple implicational invariants is to allow finite conjunctions $A_1 \land \ldots \land A_n$ of atomic formulas $A_1, \ldots, A_n \in \mathbb{A}$ as antecedents. This generalisation, though graphically not generally tractable in two dimensions, is interesting because with it, signed personality factors, represented by atomic formulas, can be analysed in terms of their *individually necessary and jointly sufficient conditions,* and thus *logically characterised in terms of each other.* More precisely, we mean by this characterisation that from

1. $\vdash_{\mathcal{I}(P)} (A_1 \land \ldots \land A_n) \to A'$, that is, the atomic formulas $A_1, \ldots, A_n$ are *jointly sufficient conditions* for the atomic formula A', and

2. $\vdash_{\mathcal{I}(P)} (A' \to A_1) \land \ldots \land (A' \to A_n)$, that is, the atomic formulas $A_1, \ldots, A_n$ are *individually necessary conditions* for the atomic formula A',

we can deduce the following *equivalence characterisation* of A' :

$$\vdash_{\mathcal{I}(P)} (A_1 \land \ldots \land A_n) \leftrightarrow A'.$$

Obviously, the truth of Item 2 can be ascertained graphically with our automatic procedure, and Item 1 can be ascertained interactively with the following semi-automatic procedure, involving a standard, efficient database query language:

1. Transcribe the given P (e.g., Table 2) into *non-recursive Datalog* [1];
2. Formulate and then query the resulting database with the jointly sufficient conditions that you suspect to be true.

Proposition 1 (Suffix closure of $\mathcal{I}$). *For every* $P, P' \in \mathrm{SPP}^{+}$*,*

1. $\mathcal{I}(P \star P') \subseteq \mathcal{I}(P')$

2. $P \sqsubseteq P'$ *implies* $\mathcal{I}(P) \subseteq \mathcal{I}(P')$

Proof. For (1), consider:

1.	$P, P' \in \mathrm{SPP}^{+}$	hypothesis
2.	$\phi \in \mathcal{I}(P \star P')$	hypothesis
3.	there are $A, A' \in \mathbb{A}$ such that $\phi = A \to A'$ and for every $P'' \in \mathrm{SPP}^{+}$, $P \star P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	2
4.	$\phi = A \to A'$ and for every $P'' \in \mathrm{SPP}^{+}$, $P \star P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	hypothesis
5.	$P'' \in \mathrm{SPP}^{+}$	hypothesis
6.	$P' \sqsubseteq P''$	hypothesis
7.	$P \star P' \sqsubseteq P''$	6, Fact 1
8.	$\mathrm{p}(\pi_1(P'')) \vdash \phi$	4, 5, 7
9.	$P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	6–8
10.	for every $P'' \in \mathrm{SPP}^{+}$, $P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	5–9
11.	$\phi = A \to A'$ and for every $P'' \in \mathrm{SPP}^{+}$, $P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	4, 10
12.	there are $A, A' \in \mathbb{A}$ such that $\phi = A \to A'$ and for every $P'' \in \mathrm{SPP}^{+}$, $P' \sqsubseteq P''$ implies $\mathrm{p}(\pi_1(P'')) \vdash \phi$	11
13.	$\phi \in \mathcal{I}(P')$	12
14.	$\phi \in \mathcal{I}(P')$	3, 4–13
15.	$\mathcal{I}(P \star P') \subseteq \mathcal{I}(P')$	2–14

16. for every $P, P' \in \mathrm{SPP}^{+}$, $\mathcal{I}(P \star P') \subseteq \mathcal{I}(P')$ 1–15.

For (2), consider:

1.	$P, P' \in \mathrm{SPP}^{+}$	hypothesis
2.	$P \sqsubseteq P'$	hypothesis
3.	$P = P'$ or there is $P'' \in \mathrm{SPP}^{+}$ such that $P = P'' \star P'$	2
4.	$P = P'$ implies $\{P\}^{\lhd} \subseteq \{P'\}^{\lhd}$	equality law
5.	there is $P'' \in \mathrm{SPP}^{+}$ such that $P = P'' \star P'$	hypothesis
6.	$P'' \in \mathrm{SPP}^{+}$ and $P = P'' \star P'$	hypothesis
7.	$\phi \in \mathcal{I}(P)$	hypothesis
8.	$\phi \in \mathcal{I}(P'' \star P')$	6, 7
9.	$\mathcal{I}(P'' \star P') \subseteq \mathcal{I}(P')$	1, Propostion 1.1
10.	$\phi \in \mathcal{I}(P')$	8, 9
11.	$\mathcal{I}(P) \subseteq \mathcal{I}(P')$	7–10
12.	$\mathcal{I}(P) \subseteq \mathcal{I}(P')$	5, 6–11
13.	there is $P'' \in \mathrm{SPP}^{+}$ such that $P = P'' \star P'$ implies $\mathcal{I}(P) \subseteq \mathcal{I}(P')$	5–12
14.	$\mathcal{I}(P) \subseteq \mathcal{I}(P')$	3, 4, 13
15.	$P \sqsubseteq P'$ implies $\mathcal{I}(P) \subseteq \mathcal{I}(P')$	2–14
16.	for every $P, P' \in \mathrm{SPP}^{+}$, $P \sqsubseteq P'$ implies $\mathcal{I}(P) \subseteq \mathcal{I}(P')$	1–15.

□

Proposition 2.

1. *$\vdash_{\mathcal{I}(P \star P')} \phi$ implies $\vdash_{\mathcal{I}(P')} \phi$*
2. *If $P \sqsubseteq P'$ and $\vdash_{\mathcal{I}(P)} \phi$ then $\vdash_{\mathcal{I}(P')} \phi$.*

Proof. Combine Fact 2.2 with Proposition 1.1 and Proposition 1.2, respectively. □

The following property means that our personality theories have the desired prime-filter property (see Proposition 3), as announced in the introduction.

Theorem 1 (Disjunction Property)**.**

$$\textit{If } \vdash_{\mathcal{I}(P)} \phi \vee \phi' \textit{ then } \vdash_{\mathcal{I}(P)} \phi \textit{ or } \vdash_{\mathcal{I}(P)} \phi'.$$

Proof. Our proof strategy is to adapt de Jongh's strategy in [5] to our simpler setting, thanks to which our proof reduces to Gödel's proof of the disjunction property of a basic intuitionistic theory [8] such as our $\mathrm{Cl}(\emptyset)$: So suppose that $\vdash_{\mathcal{I}(P)} \phi \vee \phi'$. Adapting an observation from [5], we can assert that $\vdash_{\mathcal{I}(P)} \phi \vee \phi'$ if and only if $\vdash_{\emptyset} \bigwedge\mathcal{I}(P) \to (\phi \vee \phi')$. Thus $\vdash_{\emptyset} \bigwedge\mathcal{I}(P) \to (\phi \vee \phi')$. Hence $\vdash_{\emptyset} (\bigwedge\mathcal{I}(P) \to \phi) \vee (\bigwedge\mathcal{I}(P) \to \phi')$. Hence $\vdash_{\emptyset} (\bigwedge\mathcal{I}(P) \to \phi)$ or $\vdash_{\emptyset} (\bigwedge\mathcal{I}(P) \to \phi')$ by Gödel's proof. Hence $\vdash_{\mathcal{I}(P)} \phi$ or $\vdash_{\mathcal{I}(P)} \phi'$ again by de Jongh's observation. □

3 Personality categorisation

In this section, we present the part of our framework for the mathematical categorisation of personality axioms into formal personality theories, as these axioms might have been discovered with the methodology presented in the previous section. As announced in the introduction, personality theories and personality-test data are related by a Galois-connection [4, Chapter 7]. We start with defining this connection and the two personality (powerset) spaces that it connects.

Definition 4 (Personality algebras)**.** Let the mappings ${}^{\rhd} : 2^{\mathcal{L}(\mathbb{A})} \to 2^{\mathrm{SPP}^+}$, called *right polarity,* and ${}^{\lhd} : 2^{\mathrm{SPP}^+} \to 2^{\mathcal{L}(\mathbb{A})}$, called *left polarity,* be such that

- $\Phi^{\rhd} := \{\, P \in \mathrm{SPP}^+ \mid \text{for every } \phi \in \Phi, \vdash_{\mathcal{I}(P)} \phi \,\}$ and
- $\mathcal{P}^{\lhd} := \{\, \phi \in \mathcal{L}(\mathbb{A}) \mid \text{for every } P \in \mathcal{P}, \vdash_{\mathcal{I}(P)} \phi \,\}$.

Further let $\equiv \subseteq 2^{\mathcal{L}(\mathbb{A})} \times 2^{\mathcal{L}(\mathbb{A})}$ and $\equiv \subseteq 2^{\mathrm{SPP}^+} \times 2^{\mathrm{SPP}^+}$ be their *kernels,* that is, for every $\Phi, \Phi' \in 2^{\mathcal{L}(\mathbb{A})}$, $\Phi \equiv \Phi'$ by definition if and only if $\Phi^{\rhd} = \Phi'^{\rhd}$ and for every $\mathcal{P}, \mathcal{P}' \in 2^{\mathrm{SPP}^+}$, $\mathcal{P} \equiv \mathcal{P}'$ by definition if and only if $\mathcal{P}^{\lhd} = \mathcal{P}'^{\lhd}$, respectively.

Then, for each one of the two (inclusion-ordered, Boolean) powerset algebras

$$\langle\, 2^{\mathcal{L}(\mathbb{A})}, \emptyset, \cap, \cup, \mathcal{L}(\mathbb{A}), \bar{\cdot}, \subseteq \,\rangle \underset{\lhd}{\overset{\rhd}{\rightleftharpoons}} \langle\, 2^{\mathrm{SPP}^+}, \emptyset, \cap, \cup, \mathrm{SPP}^+, \bar{\cdot}, \subseteq \,\rangle\,,$$

define its *(ordered) quotient join semi-lattice with bottom* (and thus idempotent commutative monoid) modulo its kernel as in Table 5.

Note that our focus is on the powerset and not on the quotient algebras. The purpose of the quotient algebras is simply to indicate the maximally definable algebraic structure in our context. As a matter of fact, only the join- but not the meet-operation is well-defined in the quotient algebra (see Corollary 1).

Statements	Test results
$\top \;:=\; [\mathcal{L}(\mathbb{A})]_\equiv$	$\top \;:=\; [\mathrm{SPP}^{+}]_\equiv$
$[\Phi]_\equiv \sqcup [\Phi']_\equiv \;:=\; [\Phi \cup \Phi']_\equiv$	$[\mathcal{P}]_\equiv \sqcup [\mathcal{P}']_\equiv \;:=\; [\mathcal{P} \cup \mathcal{P}']_\equiv$
$\bot \;:=\; [\emptyset]_\equiv$	$\bot \;:=\; [\emptyset]_\equiv$
$[\Phi]_\equiv \sqsubseteq [\Phi']_\equiv \;:\!\text{iff}\; [\Phi]_\equiv \sqcup [\Phi']_\equiv = [\Phi']_\equiv$	$[\mathcal{P}]_\equiv \sqsubseteq [\mathcal{P}']_\equiv \;:\!\text{iff}\; [\mathcal{P}]_\equiv \sqcup [\mathcal{P}']_\equiv = [\mathcal{P}']_\equiv$

Table 5: Quotient algebras

Proposition 3 (Basic properties of personality theories)**.**

1. $\{P\}^{\triangleleft} = (\mathrm{Cl} \circ \mathcal{I})(P)$ *(generalisation to sets)*
2. $P \sqsubseteq P'$ *implies* $\{P\}^{\triangleleft} \subseteq \{P'\}^{\triangleleft}$ *(monotonicity)*
3. *prime filter properties:*
   (a) *if* $\phi \in \{P\}^{\triangleleft}$ *and* $\phi' \in \{P\}^{\triangleleft}$ *then* $\phi \wedge \phi' \in \{P\}^{\triangleleft}$ *(and vice versa)*
   (b) *if* $\phi \in \{P\}^{\triangleleft}$ *and* $\phi' \in \mathcal{L}(\mathbb{A})$ *and* $\phi \vdash_{\mathcal{I}(P)} \phi'$ *then* $\phi' \in \{P\}^{\triangleleft}$
   (c) *if* $\phi \vee \phi' \in \{P\}^{\triangleleft}$ *then* $\phi \in \{P\}^{\triangleleft}$ *or* $\phi' \in \{P\}^{\triangleleft}$ *(and vice versa)*

   ($\{P\}^{\triangleleft}$ *is an intuitionistic theory.)*
4. *for every* $\phi, \phi', \phi'' \in \mathcal{L}(\mathbb{A})$,
$$\textit{if } \phi \vee \phi', \phi \vee \phi'' \in \bigcap_{P \in \mathcal{P}} \{P\}^{\triangleleft} \textit{ then } \phi \vee (\phi' \wedge \phi'') \in \bigcap_{P \in \mathcal{P}} \{P\}^{\triangleleft}$$
   ($\bigcap_{P \in \mathcal{P}} \{P\}^{\triangleleft}$ *is a distributive filter.)*

Proof. For (1), consider that

$$\begin{aligned}
\{P\}^{\triangleleft} &= \{\, \phi \in \mathcal{L}(\mathbb{A}) \mid \text{for every } P' \in \{P\}, \vdash_{\mathcal{I}(P')} \phi \,\} \\
&= \{\, \phi \in \mathcal{L}(\mathbb{A}) \mid \vdash_{\mathcal{I}(P)} \phi \,\} \\
&= \{\, \phi \in \mathcal{L}(\mathbb{A}) \mid \phi \in \mathrm{Cl}(\mathcal{I}(P)) \,\} \\
&= \mathrm{Cl}(\mathcal{I}(P)) \\
&= (\mathrm{Cl} \circ \mathcal{I})(P)
\end{aligned}$$

For (2), suppose that $P \sqsubseteq P'$. Hence $\mathcal{I}(P) \subseteq \mathcal{I}(P')$ by Proposition 1.2. Hence $\mathrm{Cl}(\mathcal{I}(P)) \subseteq \mathrm{Cl}(\mathcal{I}(P'))$ by Fact 2.2. Thus $\{P\}^{\triangleleft} \subseteq \{P'\}^{\triangleleft}$ by (1). (3.a) follows from the fact that $(\phi \to (\phi' \to (\phi \land \phi')) \in \mathrm{Cl}(\emptyset)$ (and $((\phi \land \phi') \to \phi), ((\phi \land \phi') \to \phi') \in \mathrm{Cl}(\emptyset)$), Fact 2.2, and (1). For (3.b), inspect definitions, and for (3.c), Theorem 1, definitions, and (1). For (4), consider (3) and recall that intersections of prime filters are distributive filters [4, Exercise 10.9]. □

Now note the two macro-definitions ${}^{\rhd\lhd} := {}^{\rhd} \circ {}^{\lhd}$ and ${}^{\lhd\rhd} := {}^{\lhd} \circ {}^{\rhd}$ with $\circ$ being function composition, as usual (from right to left, as usual too).

Lemma 1 (Some useful properties of ${}^{\rhd}$ and ${}^{\lhd}$)**.**

1. *if* $\Phi \subseteq \Phi'$ *then* $\Phi'^{\rhd} \subseteq \Phi^{\rhd}$ *(${}^{\rhd}$ is antitone)*
2. *if* $\mathcal{P} \subseteq \mathcal{P}'$ *then* $\mathcal{P}'^{\lhd} \subseteq \mathcal{P}^{\lhd}$ *(${}^{\lhd}$ is antitone)*
3. $\mathcal{P} \subseteq (\mathcal{P}^{\lhd})^{\rhd}$ *(${}^{\rhd\lhd}$ is extensive)*
4. $\Phi \subseteq (\Phi^{\rhd})^{\lhd}$ *(${}^{\lhd\rhd}$ is extensive)*
5. $((\mathcal{P}^{\lhd})^{\rhd})^{\lhd} = \mathcal{P}^{\lhd}$
6. $((\Phi^{\rhd})^{\lhd})^{\rhd} = \Phi^{\rhd}$
7. $(((\mathcal{P}^{\lhd})^{\rhd})^{\lhd})^{\rhd} = (\mathcal{P}^{\lhd})^{\rhd}$ *(${}^{\rhd\lhd}$ is idempotent)*
8. $(((\Phi^{\rhd})^{\lhd})^{\rhd})^{\lhd} = (\Phi^{\rhd})^{\lhd}$ *(${}^{\lhd\rhd}$ is idempotent)*
9. *if* $\mathcal{P} \subseteq \mathcal{P}'$ *then* $(\mathcal{P}^{\lhd})^{\rhd} \subseteq (\mathcal{P}^{\lhd})^{\rhd}$ *(${}^{\rhd\lhd}$ is monotone)*
10. *if* $\Phi \subseteq \Phi'$ *then* $(\Phi^{\rhd})^{\lhd} \subseteq (\Phi'^{\rhd})^{\lhd}$ *(${}^{\lhd\rhd}$ is monotone)*

Proof. For (1), suppose that $\Phi \subseteq \Phi'$. Further suppose that $P \in \Phi'^{\rhd}$. That is, $\Phi' \subseteq \mathrm{Cl}(\mathcal{I}(P))$. Now suppose that $\phi \in \Phi$. Hence $\phi \in \Phi'$. Hence $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Thus $\Phi \subseteq \mathrm{Cl}(\mathcal{I}(P))$. That is, $P \in \Phi^{\rhd}$. Thus $\Phi'^{\rhd} \subseteq \Phi^{\rhd}$. For (2), suppose that $\mathcal{P} \subseteq \mathcal{P}'$. Further suppose that $\phi \in \mathcal{P}'^{\lhd}$. That is, for every $P \in \mathcal{P}'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Now suppose that $P \in \mathcal{P}$. Hence $P \in \mathcal{P}'$. Hence $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Thus for every $P \in \mathcal{P}$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. That is, $\phi \in \mathcal{P}^{\lhd}$. Thus $\mathcal{P}'^{\lhd} \subseteq \mathcal{P}^{\lhd}$. For (3), suppose that $P \in \mathcal{P}$. Further suppose that $\phi \in \mathcal{P}^{\lhd}$. That is, for every $P \in \mathcal{P}$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Hence $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Thus $\mathcal{P}^{\lhd} \subseteq \mathrm{Cl}(\mathcal{I}(P))$. That is, $P \in (\mathcal{P}^{\lhd})^{\rhd}$. Thus $\mathcal{P} \subseteq (\mathcal{P}^{\lhd})^{\rhd}$. For (4), suppose that $\phi \in \Phi$. Further suppose that $P \in \Phi^{\rhd}$. That is, $\Phi \subseteq \mathrm{Cl}(\mathcal{I}(P))$. Hence $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. Thus for every $P \in \Phi^{\rhd}$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$. That is, $\phi \in (\Phi^{\rhd})^{\lhd}$. Thus $\Phi \subseteq (\Phi^{\rhd})^{\lhd}$. For (5), consider that $\mathcal{P}^{\lhd} \subseteq ((\mathcal{P}^{\lhd})^{\rhd})^{\lhd}$ is an instance of (4), and that $((\mathcal{P}^{\lhd})^{\rhd})^{\lhd} \subseteq \mathcal{P}^{\lhd}$ by the application of (2) to (3). For (6), consider that $\Phi^{\rhd} \subseteq ((\Phi^{\rhd})^{\lhd})^{\rhd}$ is an instance of (3), and that $((\Phi^{\rhd})^{\lhd})^{\rhd} \subseteq \Phi^{\rhd}$ by the application of (1) to (4). For (7) and (8), substitute $P^{\lhd}$ for Φ in (6), and $\Phi^{\rhd}$ for $\mathcal{P}$ in (5), respectively. For (9) and (10), transitively apply (1) to (2) and (2) to (1), respectively. □

Notice that Lemma 1.3, 1.7, and 1.9 together with Lemma 1.4, 1.8, and 1.10 mean that ${}^{\triangleleft\triangleright}$ and ${}^{\triangleright\triangleleft}$ are closure operators, which is a fact relevant to Theorem 3.

Theorem 2 (The Galois-connection property of $({}^{\triangleright}, {}^{\triangleleft})$)**.** *The ordered pair* $({}^{\triangleright}, {}^{\triangleleft})$ *is an* antitone *or* order-reversing Galois-connection *between the powerset algebras in Definition 4. That is, for every* $\Phi \in 2^{\mathcal{L}(\mathbb{A})}$ *and* $\mathcal{P} \in 2^{\mathrm{SPP}^+}$,

$$\mathcal{P} \subseteq \Phi^{\triangleright} \textit{ if and only if } \Phi \subseteq \mathcal{P}^{\triangleleft}.$$

Proof. Let $\Phi \in 2^{\mathcal{L}(\mathbb{A})}$ and $\mathcal{P} \in 2^{\mathrm{SPP}^+}$ and suppose that $\mathcal{P} \subseteq \Phi^{\triangleright}$. Hence $(\Phi^{\triangleright})^{\triangleleft} \subseteq \mathcal{P}^{\triangleleft}$ by Lemma 1.2. Further, $\Phi \subseteq (\Phi^{\triangleright})^{\triangleleft}$ by Lemma 1.4. Hence $\Phi \subseteq \mathcal{P}^{\triangleleft}$ by transitivity. Conversely suppose that $\Phi \subseteq \mathcal{P}^{\triangleleft}$. Hence $(\mathcal{P}^{\triangleleft})^{\triangleright} \subseteq \Phi^{\triangleright}$ by Lemma 1.1. Further, $\mathcal{P} \subseteq (\mathcal{P}^{\triangleleft})^{\triangleright}$ by Lemma 1.3. Hence $\mathcal{P} \subseteq \Phi^{\triangleright}$. □

Galois-connections are connected to *residuated mappings* [2].

Theorem 3 (De-Morgan like laws)**.**

1. $$\begin{aligned}(\mathcal{P} \cup \mathcal{P}')^{\triangleleft} = \mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft} = ((\mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft})^{\triangleright})^{\triangleleft} &\subseteq \\ \mathcal{P}^{\triangleleft} \cup \mathcal{P}'^{\triangleleft} \subseteq ((\mathcal{P}^{\triangleleft} \cup \mathcal{P}'^{\triangleleft})^{\triangleright})^{\triangleleft} &\subseteq (\mathcal{P} \cap \mathcal{P}')^{\triangleleft}\end{aligned}$$

2. $$\begin{aligned}(\Phi \cup \Phi')^{\triangleright} = \Phi^{\triangleright} \cap \Phi'^{\triangleright} = ((\Phi^{\triangleright} \cap \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} &\subseteq \\ \Phi^{\triangleright} \cup \Phi'^{\triangleright} \subseteq ((\Phi^{\triangleright} \cup \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} &\subseteq (\Phi \cap \Phi')^{\triangleright}\end{aligned}$$

Proof. For $(\mathcal{P} \cup \mathcal{P}')^{\triangleleft} = \mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft}$ (join becomes meet) in (1), let $\phi \in \mathcal{L}(\mathbb{A})$, and consider that $\phi \in (\mathcal{P} \cup \mathcal{P}')^{\triangleleft}$ if and only if (for every $P \in \mathcal{P} \cup \mathcal{P}'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) if and only if [for every P, ($P \in \mathcal{P}$ or $P \in \mathcal{P}'$) implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$] if and only if [for every P, ($P \in \mathcal{P}$ implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) and ($P \in \mathcal{P}'$ implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$)] if and only if [(for every $P \in \mathcal{P}$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) and (for every $P \in \mathcal{P}'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$)] if and only if ($\phi \in \mathcal{P}^{\triangleleft}$ and $\phi \in \mathcal{P}'^{\triangleleft}$) if and only if $\phi \in \mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft}$. Then, $\mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft} \subseteq \mathcal{P}^{\triangleleft} \cup \mathcal{P}'^{\triangleleft}$ by elementary set theory. For later use of $\mathcal{P}^{\triangleleft} \cup \mathcal{P}'^{\triangleleft} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\triangleleft}$ in (1) consider:

1.	$\phi \in \mathcal{P}^{\triangleleft} \cup \mathcal{P}'^{\triangleleft}$	hypothesis
2.	$\phi \in \mathcal{P}^{\triangleleft}$ or $\phi \in \mathcal{P}'^{\triangleleft}$	1
3.	$\quad \phi \in \mathcal{P}^{\triangleleft}$	hypothesis
4.	$\quad\quad P \in \mathcal{P} \cap \mathcal{P}'$	hypothesis
5.	$\quad\quad P \in \mathcal{P}$ and $P \in \mathcal{P}'$	4
6.	$\quad\quad P \in \mathcal{P}$	5
7.	$\quad\quad \{P\} \subseteq \mathcal{P}$	6
8.	$\quad\quad \mathcal{P}^{\triangleleft} \subseteq \{P\}^{\triangleleft}$	7, Lemma 1.2
9.	$\quad\quad \phi \in \{P\}^{\triangleleft}$	3, 8

10.	$\qquad\phi \in \mathrm{Cl}(\mathcal{I}(P))$	9
11.	$\quad$ for every $P \in \mathcal{P} \cap \mathcal{P}'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$	4–10
12.	$\quad\phi \in (\mathcal{P} \cap \mathcal{P}')^{\lhd}$	11
13.	if $\phi \in \mathcal{P}^{\lhd}$ then $\phi \in (\mathcal{P} \cap \mathcal{P}')^{\lhd}$	3–12
14.	if $\phi \in \mathcal{P}'^{\lhd}$ then $\phi \in (\mathcal{P} \cap \mathcal{P}')^{\lhd}$	similarly to 3–12 for 13
15.	$\phi \in (\mathcal{P} \cap \mathcal{P}')^{\lhd}$	2, 13, 14
16.	$\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\lhd}$	1–15.

For $((\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd})^{\rhd})^{\lhd} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\lhd}$ in (1), consider the previously proved property that $\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\lhd}$. Hence $(\mathcal{P} \cap \mathcal{P}') \subseteq (\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd})^{\rhd}$ by Theorem 2. Hence $((\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd})^{\rhd})^{\lhd} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\lhd}$ by Lemma 1.2. Then, $\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd} \subseteq ((\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd})^{\rhd})^{\lhd}$ in (1) is an instance of Lemma 1.4. For $\mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd} = ((\mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd})^{\rhd})^{\lhd}$ in (1), consider that $((\mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd})^{\rhd})^{\lhd} = (((\mathcal{P} \cup \mathcal{P}')^{\lhd})^{\rhd})^{\lhd}$ by the previously proved property that $\mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd} = (\mathcal{P} \cup \mathcal{P}')^{\lhd}$. But $(((\mathcal{P} \cup \mathcal{P}')^{\lhd})^{\rhd})^{\lhd} = (\mathcal{P} \cup \mathcal{P}')^{\lhd}$ by Lemma 1.5. Hence $((\mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd})^{\rhd})^{\lhd} = \mathcal{P}^{\lhd} \cap \mathcal{P}'^{\lhd}$.

For $(\Phi \cup \Phi')^{\rhd} = \Phi^{\rhd} \cap \Phi'^{\rhd}$ (join becomes meet) in (2), let $P \in \mathrm{SPP}^{+}$, and consider that $P \in (\Phi \cup \Phi')^{\rhd}$ if and only if (for every $\phi \in \Phi \cup \Phi'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) if and only if [for every ϕ, ($\phi \in \Phi$ or $\phi \in \Phi'$) implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$] if and only if [for every ϕ, ($\phi \in \Phi$ implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) and ($\phi \in \Phi'$ implies $\phi \in \mathrm{Cl}(\mathcal{I}(P))$)] if and only if [(for every $\phi \in \Phi$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$) and (for every $\phi \in \Phi'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$)] if and only if ($P \in \Phi^{\rhd}$ and $P \in \Phi'^{\rhd}$) if and only if $P \in \Phi^{\rhd} \cap \Phi'^{\rhd}$. Then, $\Phi^{\rhd} \cap \Phi'^{\rhd} \subseteq \Phi^{\rhd} \cup \Phi'^{\rhd}$ by elementary set theory. For later use of $\Phi^{\rhd} \cup \Phi'^{\rhd} \subseteq (\Phi \cap \Phi')^{\rhd}$ in (2) consider:

1.	$P \in \Phi^{\rhd} \cup \Phi'^{\rhd}$	hypothesis
2.	$P \in \Phi^{\rhd}$ or $P \in \Phi'^{\rhd}$	1
3.	$\quad P \in \Phi^{\rhd}$	hypothesis
4.	$\qquad\phi \in \Phi \cap \Phi'$	hypothesis
5.	$\qquad\phi \in \Phi$ and $\phi \in \Phi'$	4
6.	$\qquad\phi \in \Phi$	5
7.	$\qquad\{\phi\} \subseteq \Phi$	6
8.	$\qquad\Phi^{\rhd} \subseteq \{\phi\}^{\rhd}$	7, Lemma 1.1
9.	$\qquad P \in \{\phi\}^{\rhd}$	3, 8
10.	$\qquad\phi \in \mathrm{Cl}(\mathcal{I}(P))$	9
11.	$\quad$ for every $\phi \in \Phi \cap \Phi'$, $\phi \in \mathrm{Cl}(\mathcal{I}(P))$	4–10
12.	$\quad P \in (\Phi \cap \Phi')^{\rhd}$	11
13.	if $P \in \Phi^{\rhd}$ then $P \in (\Phi \cap \Phi')^{\rhd}$	3–12
14.	if $P \in \Phi'^{\rhd}$ then $P \in (\Phi \cap \Phi')^{\rhd}$	similarly to 3–12 for 13

15. $\quad P \in (\Phi \cap \Phi')^{\triangleright}$ 2, 13, 14

16. $\Phi^{\triangleright} \cup \Phi'^{\triangleright} \subseteq (\Phi \cap \Phi')^{\triangleright}$ 1–15.

For $((\Phi^{\triangleright} \cup \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} \subseteq (\Phi \cap \Phi')^{\triangleright}$ in (2), consider the previously proved property that $\Phi^{\triangleright} \cup \Phi'^{\triangleright} \subseteq (\Phi \cap \Phi')^{\triangleright}$. Hence $(\Phi \cap \Phi') \subseteq (\Phi^{\triangleright} \cup \Phi'^{\triangleright})^{\triangleleft}$ by Theorem 2. Hence $((\Phi^{\triangleright} \cup \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} \subseteq (\Phi \cap \Phi')^{\triangleright}$ by Lemma 1.1. Then, $\Phi^{\triangleright} \cup \Phi'^{\triangleright} \subseteq ((\Phi^{\triangleright} \cup \Phi'^{\triangleright})^{\triangleleft})^{\triangleright}$ in (2) is an instance of Lemma 1.3. For $\Phi^{\triangleright} \cap \Phi'^{\triangleright} = ((\Phi^{\triangleright} \cap \Phi'^{\triangleright})^{\triangleleft})^{\triangleright}$ in (2), consider that $((\Phi^{\triangleright} \cap \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} = (((\Phi \cup \Phi')^{\triangleright})^{\triangleleft})^{\triangleright}$ by the previously proved property that $\Phi^{\triangleright} \cap \Phi'^{\triangleright} = (\Phi \cup \Phi')^{\triangleright}$. But $(((\Phi \cup \Phi')^{\triangleright})^{\triangleleft})^{\triangleright} = (\Phi \cup \Phi')^{\triangleright}$ by Lemma 1.6. Hence $((\Phi^{\triangleright} \cap \Phi'^{\triangleright})^{\triangleleft})^{\triangleright} = \Phi^{\triangleright} \cap \Phi'^{\triangleright}$. □

Corollary 1. *The quotient algebras in Table 5 are well-defined, that is, the equivalence relations* $\equiv \subseteq 2^{\mathcal{L}(\mathbb{A})} \times 2^{\mathcal{L}(\mathbb{A})}$ *and* $\equiv \subseteq 2^{\mathrm{SPP}^+} \times 2^{\mathrm{SPP}^+}$ *are* congruences:

1. *if* $\Phi \equiv \Phi'$ *and* $\Phi'' \equiv \Phi'''$ *then* $\Phi \cup \Phi'' \equiv \Phi' \cup \Phi'''$*;*

2. *if* $\mathcal{P} \equiv \mathcal{P}'$ *and* $\mathcal{P}'' \equiv \mathcal{P}'''$ *then* $\mathcal{P} \cup \mathcal{P}'' \equiv \mathcal{P}' \cup \mathcal{P}'''$*.*

Proof. By the De-Morgan like laws $(\Phi \cup \Phi')^{\triangleright} = \Phi^{\triangleright} \cap \Phi'^{\triangleright}$ and $(\mathcal{P} \cup \mathcal{P}')^{\triangleleft} = \mathcal{P}^{\triangleleft} \cap \mathcal{P}'^{\triangleleft}$, respectively (see Theorem 3). □

We are finally ready for defining our announced personality categories, and this by means of our previously-defined Galois-connection.

Definition 5 (Personality categories)**.** Let $\mathcal{P} \in 2^{\mathrm{SPP}^+}$ and $\Phi \in 2^{\mathcal{L}(\mathbb{A})}$, and let

- $\mathcal{T}_{\Phi} := \{\, \tau : 2^{\mathrm{SPP}^+} \to 2^{\mathrm{SPP}^+} \mid \begin{array}{c} \text{for every } \mathcal{P} \in 2^{\mathrm{SPP}^+} \text{ and } \phi \in \Phi, \\ \phi \in \mathcal{P}^{\triangleleft} \text{ implies } \phi \in \tau(\mathcal{P})^{\triangleleft} \end{array} \,\}$ and

- $\mathcal{T}_{\mathcal{P}} := \{\, \tau : 2^{\mathcal{L}(\mathbb{A})} \to 2^{\mathcal{L}(\mathbb{A})} \mid \begin{array}{c} \text{for every } \Phi \in 2^{\mathcal{L}(\mathbb{A})} \text{ and } P \in \mathcal{P}, \\ P \in \Phi^{\triangleright} \text{ implies } P \in \tau(\Phi)^{\triangleright} \end{array} \,\}$.

Then, define the categories (monoids)

$$\mathbf{T}_{\Phi} := \langle \mathcal{T}_{\Phi}, \mathrm{id}, \circ \rangle \text{ and } \mathbf{T}_{\mathcal{P}} := \langle \mathcal{T}_{\mathcal{P}}, \mathrm{id}, \circ \rangle$$

of Φ- and $\mathcal{P}$-preserving transformations, respectively.

Proposition 4 (Antitonicity properties of personality categories)**.**

1. $\Phi \subseteq \Phi'$ *implies* $\mathcal{T}_{\Phi'} \subseteq \mathcal{T}_{\Phi}$

2. $\mathcal{P} \subseteq \mathcal{P}'$ *implies* $\mathcal{T}_{\mathcal{P}'} \subseteq \mathcal{T}_{\mathcal{P}}$

3. $P \sqsubseteq P'$ *implies* $\mathcal{T}_{\{P'\}^{\lhd}} \subseteq \mathcal{T}_{\{P\}^{\lhd}}$

4. $\mathcal{T}_{\Phi \cup \Phi'} \subseteq \mathcal{T}_{\Phi} \subseteq \mathcal{T}_{\Phi \cap \Phi'}$

5. $\mathcal{T}_{\mathcal{P} \cup \mathcal{P}'} \subseteq \mathcal{T}_{\mathcal{P}} \subseteq \mathcal{T}_{\mathcal{P} \cap \mathcal{P}'}$

Proof. (1) and (2) follow straightforwardly from their respective definition, and (4) and (5) from (1) and (2), respectively. (3) follows from Proposition 3.2 and (1) by transitivity. □

Proposition 5 (Preservation properties of personality transformations)**.**

1. $\tau \in \mathcal{T}_{\Phi^{\rhd}}$ *implies* $\Phi^{\rhd} \subseteq \tau(\Phi)^{\rhd}$

2. $\tau \in \mathcal{T}_{\mathcal{P}^{\lhd}}$ *implies* $\mathcal{P}^{\lhd} \subseteq \tau(\mathcal{P})^{\lhd}$

3. $\tau \in \mathcal{T}_{(\Phi \cap \Phi')^{\rhd}}$ *implies (*$\Phi^{\rhd} \subseteq \tau(\Phi)^{\rhd}$ *and* $\Phi'^{\rhd} \subseteq \tau(\Phi')^{\rhd}$*)*

4. $\tau \in \mathcal{T}_{(\mathcal{P} \cap \mathcal{P}')^{\lhd}}$ *implies (*$\mathcal{P}^{\lhd} \subseteq \tau(\mathcal{P})^{\lhd}$ *and* $\mathcal{P}'^{\lhd} \subseteq \tau(\mathcal{P}')^{\lhd}$*)*

Proof. (1) and (2) follow by expansion of definitions. For (3) suppose that $\tau \in \mathcal{T}_{(\Phi \cap \Phi')^{\rhd}}$. But by Theorem 3.2, $\Phi^{\rhd} \cup \Phi'^{\rhd} \subseteq (\Phi \cap \Phi')^{\rhd}$. Hence $\mathcal{T}_{(\Phi \cap \Phi')^{\rhd}} \subseteq \mathcal{T}_{\Phi^{\rhd} \cup \Phi'^{\rhd}}$ by Proposition 4.1. Hence $\tau \in \mathcal{T}_{\Phi^{\rhd} \cup \Phi'^{\rhd}}$. Hence $\tau \in \mathcal{T}_{\Phi^{\rhd}}$ and $\tau \in \mathcal{T}_{\Phi'^{\rhd}}$ by Proposition 4.4. Hence $\Phi^{\rhd} \subseteq \tau(\Phi)^{\rhd}$ and $\Phi'^{\rhd} \subseteq \tau(\Phi')^{\rhd}$ by (1). For (4) suppose that $\tau \in \mathcal{T}_{(\mathcal{P} \cap \mathcal{P}')^{\lhd}}$. But by Theorem 3.1, $\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd} \subseteq (\mathcal{P} \cap \mathcal{P}')^{\lhd}$. Hence $\mathcal{T}_{(\mathcal{P} \cap \mathcal{P}')^{\lhd}} \subseteq \mathcal{T}_{\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd}}$ by Proposition 4.2. Hence $\tau \in \mathcal{T}_{\mathcal{P}^{\lhd} \cup \mathcal{P}'^{\lhd}}$. Hence $\tau \in \mathcal{T}_{\mathcal{P}^{\lhd}}$ and $\tau \in \mathcal{T}_{\mathcal{P}'^{\lhd}}$ by Proposition 4.5. Hence $\mathcal{P}^{\lhd} \subseteq \tau(\mathcal{P})^{\lhd}$ and $\mathcal{P}'^{\lhd} \subseteq \tau(\mathcal{P}')^{\lhd}$ by (2). □

Typically, there are many invariant-preserving transformations of a person's personality as represented by their personality-test result, such as Table 2. For example, the replacement of any (∀) row j by any (∀) row i for $1 \leq i \leq j \leq 10$ in Table 2 is one (∃) such invariant-preserving transformation, as can be seen by inspecting our algorithm in Listing 1 (discounting distance to invariance). Such a replacement cannot not discount distance to invariance, and thus can never violate an invariant (not reach 0) that held before the replacement.

4 Conclusions

We have provided a formal framework for the computer-aided discovery and categorisation of personality axioms as summarised in the abstract of this paper. Our framework is meant as a contribution towards practicing psychological research with the methods of the exact sciences, for obvious ethical reasons. Psychology workers (psychologists, psychiatrists, etc.) can now apply our visual framework in their own field (of) studies in order to discover personality theories and categories of their own interest. Our hope is that these field studies will lead to a mathematical systematisation of the academic discipline of psychology in the area of test-based personality theories with the help of our framework.

As future work on our current *synchronic* data analytics approach, which infers *perfect* implicational correlations (between human reactions) at a given time point (within a Szondi personality profile, an SPP) from their invariance across time (within an SPP-sequence, a Szondi-test result), *approximate* implicational correlations can be studied and a *diachronic* data analytics approach can be taken. Actually, our implicational diagrams such as Table 3 already contain such approximate implicational correlations in the form of cell values greater than 0, which as explained on Page 13 indicate the distance to invariance and thus the approximation to the perfection in question. This notion of approximate implicational correlation can be understood and further studied as a notion of *fuzzy implication* [10]. Then, a diachronic approach would mine correlations between *different* time points, typically one or several past or present and one or several future, in order to *forecast and predict* future reactions of the person in question, such as can be done with *Bayesian inference* [19] and *time series* analysis and forecasting [20]. Actually, Table 2 is such a time series.

Last but not least, we mention the only piece of related work [3] that we are aware of. There, the author develops a framework similarly motivated by invariance as ours, but with quite different setup, outcomes, and results. The author's setup on the invariants side is a set of relations over a finite domain closed under the Boolean operations, whereas our corresponding setup is an intuitionistic theory, a certain set of propositional formulas, as induced by a data sequence (as exemplified by one produced by a personality test). On the transformations side, the author's setup is a system of injective total functions, whereas our corresponding setup is a system of total functions *tout court*.

Acknowledgements The LaTeX-packages Listings and TikZ were helpful.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases.* Addison-Wesley, 1995.

[2] T.S. Blyth. *Lattices and Ordered Algebraic Structures.* Springer, 2005.

[3] L. Burigana. Invariant relations in a finite domain. *Mathématiques et sciences humaines*, 169, 2005.

[4] B.A. Davey and H.A. Priestley. *Introduction to Lattices and Order.* Cambridge University Press, 2nd edition, 1990 (2002).

[5] D. de Jongh. The disjunction property according to Kreisel and Putnam, 2009.

[6] C.A. Furia, B. Meyer, and S. Velder. Loop invariants: Analysis, classification, and examples. *ACM Computing Surveys*, 46(3), 2014.

[7] D.M. Gabbay, editor. *What Is a Logical System?* Number 4 in Studies in Logic and Computation. Oxford University Press, 1995.

[8] K. Gödel. *Kurt Gödel: Collected Works*, volume I, chapter On the intuitionistic propositional calculus. Oxford University Press, 1986.

[9] T. Hawkings. The Erlanger programm of Felix Klein: reflections on its place in the history of mathematics. *Historia Mathematica*, 11, 1984.

[10] B. Jayaram and M. Baczyński. *Fuzzy Implications.* Springer, 2008.

[11] F. Klein. Vergleichende Betrachtungen über neuere geometrische Forschungen. *Reprinted with additional footnotes in Mathematische Annalen*, 43 (1893), 1872.

[12] F. Klein. *Elementary Mathematics from an Advanced Standpoint: Geometry.* Dover Publications, 2004.

[13] S. Kramer. Computer-aided discovery and categorisation of personality axioms. Technical Report 1403.6048, arXiv, 2014. `http://arxiv.org/abs/1403.6048`.

[14] S. Kramer. A Galois-connection between Myers-Briggs' Type Indicators and Szondi's Personality Profiles. Technical Report 1403.2000, arXiv, 2014. `http://arxiv.org/abs/1403.2000`.

[15] S. Kramer. `www.szondi-test.ch`. forthcoming.

[16] S.A. Kripke. *Formal Systems and Recursive Functions*, volume 40 of *Studies in Logic and the Foundations of Mathematics*, chapter Semantical Analysis of Intuitionistic Logic I. Elsevier, 1965.

[17] Jean-Pierre Marquis. Category theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Summer 2013 edition, 2013.

[18] J. Moschovakis. Intuitionistic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Summer 2010 edition, 2010.

[19] J. Pearl. *Causality.* Cambridge University Press, 2nd edition, 2009.

[20] J.C. Sprott. *Chaos and Time-Series Analysis.* Oxford University Press, 2003.

[21] L. Szondi. *Triebpathologie. Teil A: Dialektische Trieblehre und dialektische Methodik der Testanalyse*, volume 1. Hans Huber, 2nd edition, 1952 (1977).

[22] L. Szondi. *Lehrbuch der Experimentellen Triebdiagnostik*, volume I: Text-Band. Hans Huber, 3rd edition, 1972.

[23] L. Szondi. *Ich-Analyse: Die Grundlage zur Vereinigung der Tiefenpsychologie.* Hans Huber, 1999. English translation: `https://sites.google.com/site/ajohnstontranslationsofszondi/`.

Received March 2014