3 The BvH and HK Definitions

BvH [5] work within a game-theoretic framework and do not use causal models, so in order to compare their approach to mine we need to first translate it into the language of causal models. I do not delve into the details but rather offer a rough sketch of such a translation.

Similar to causal models, BvH represent the agents' influence on the outcome O by way of a function. Yet instead of letting some endogenous variables $\vec{A}$ represent the actions of agents directly, they use variables to represent the *strategies* that each agent can adopt to guide their actions. Aside from that, the main difference between the two formalisms is that theirs is unable to represent indirect causal relations. [3] In general, the equations of a causal model allow for an unlimited number of intermediate ancestors between variables $\vec{A}$ and an outcome variable O, so that causal influence from an agent's action can be passed on along intermediate variables to the outcome variable. BvH's outcome function on the other hand abstracts away from any mediated form of causal influence, so that the strategies causally determine the outcome *directly*. As a result, their games are to be interpreted as a single-equation causal model of the form $O = f_O(\vec{A})$. (Since the variables$\vec{A}$ are determined directly by the context $\vec{u}$, I adopt the standard practice of leaving their equations implicit.)

BvH use the famous NESS definition of causation that was proposed by Wright [32, 33] – and also formed the inspiration for the Halpern & Pearl (HP) definitions [22, 15, 16] – which states that causes are *Necessary Elements of a Sufficient Set* for the effect. Taking into account the previous remarks, it is more accurate to speak of the *direct NESS* definition. I here present my recent formalization of both the direct and the indirect NESS definitions using causal models [2]. First we need to define causal sufficiency. As do BvH, I take it to mean that a set guarantees the effect regardless of the values of the variables outside of the set.

Definition 3 (Sufficiency). *We say that* $\vec{X} = \vec{x}$ *is* sufficient *for* $Y = y$ *w.r.t.* $(M, \vec{u})$ *if* $Y \notin \vec{X}$ *and for all values* $\vec{z} \in \mathcal{R}(\vec{Z})$ *where* $\vec{Z} = \mathcal{V} - (\vec{X} \cup \{Y\})$*, it holds that* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{Z} \leftarrow \vec{z}]Y = y$.

Direct NESS-causation is then defined by stating that:

- the candidate cause and the effect actually occurred;
- the candidate cause is a member of a sufficient set;
- and it is necessary for the set to be sufficient.

Definition 4 (Direct NESS). $X = x$ directly NESS-causes $Y = y$ *w.r.t.* $(M, \vec{u})$ *if there exists a* $\vec{W} = \vec{w}$ *so that the following conditions hold:*

DN1. $(M, \vec{u}) \models X = x \wedge \vec{W} = \vec{w} \wedge Y = y$.

DN2. $X = x \wedge \vec{W} = \vec{w}$ *is sufficient for* $Y = y$ *w.r.t.* $(M, \vec{u})$.

DN3. $\vec{W} = \vec{w}$ *is not sufficient for* $Y = y$ *w.r.t.* $(M, \vec{u})$.

We can now formulate the counterpart of the BvH definition using causal models by filling in their conditions into our Responsibility Schema.

Definition 5 (BvH Responsibility). *An agent who performs* $A = a$ *is responsible for outcome* $O = o$ *w.r.t. a responsibility setting* $(M, \vec{u}, \mathcal{E})$ *if:*

- **(Causal Condition)** $A = a$ *directly NESS-causes* $O = o$ *w.r.t.* $(M, \vec{u})$.
- **(Epistemic Condition)** *There exists* $a' \in \mathcal{R}(A)$ *so that* $\Pr(A = a$ *directly NESS-causes* $O = o) > \Pr(A = a'$ *directly NESS-causes* $O = o)$.[4]

Informally, the BvH definition of responsibility requires that an agent's action directly NESS-caused the outcome, and that the agent believes they failed to minimize the probability of their action causing the outcome. The following example (taken from BvH) illustrates their definition.

[3] As I said, this is a rough sketch. Technically, one should distinguish between games in *normal form*, which is the form considered by BvH, and games in *extensive form*, from which the normal form games have been derived. Games in extensive form do allow for indirect relations, and thus there might be a way of representing indirect causal relations in game theory after all.

[4] Of course these probabilities have to be read as being conditioned on the corresponding action, i.e., as "the agent's probability that the action would cause the outcome if it were performed".