The use of partial order relations and measure theory in developing objective measures of explainability

Wim De Mulder[1,2]

[1]University of Ghent, Belgium

[2]KU Leuven, Belgium

Monday, May 9

Outline

Conflicting views on degree of explainability

Very general definition of explanation

- Explanation $E = I_1 \cup \ldots \cup I_n$
  - Information element I_j = "cognitive chunk"

(Theoretical) computer science

- First view: the more information elements, the higher the degree of explainability
- Second view: the more information elements, the lower the degree of explainability

(Experimental) social sciences

- Explanation selection: people *select* limited number of causes as *the* explanation
- Information overload

Conflicting views on degree of explainability

Very general definition of explanation

- Explanation $E = I_1 \cup \ldots \cup I_n$
  - Information element I_j = "cognitive chunk"

(Theoretical) computer science

- First view: the more information elements, the higher the degree of explainability
- Second view: the more information elements, the lower the degree of explainability

(Experimental) social sciences

- Explanation selection: people *select* limited number of causes as *the* explanation
- Information overload

Conflicting views on degree of explainability

Very general definition of explanation

- Explanation $E = I_1 \cup \ldots \cup I_n$
  - Information element I_j = "cognitive chunk"

(Theoretical) computer science

- First view: the more information elements, the higher the degree of explainability
- Second view: the more information elements, the lower the degree of explainability

(Experimental) social sciences

- Explanation selection: people *select* limited number of causes as *the* explanation
- Information overload

Conflicting views on nature of explainability

Need for an objective measure of explainability

- GDPR requires fully-automated decisions to be explainable
- This in turn requires to discern explainable from non explainable models
  - Objective minimum threshold and measure of explainability needed
  - cf. requiring citizens to drive slowly requires a specific speed limit and objective measure instrument

Need for a subjective measure of explainability

- Different categories of users require different explanations
- Mohseni, Zarei, and Ragan (2018):
  - Novice users
  - Domain experts
  - AI experts

Conflicting views on nature of explainability

Need for an objective measure of explainability

- GDPR requires fully-automated decisions to be explainable
- This in turn requires to discern explainable from non explainable models
  - Objective minimum threshold and measure of explainability needed
  - cf. requiring citizens to drive slowly requires a specific speed limit and objective measure instrument

Need for a subjective measure of explainability

- Different categories of users require different explanations
- Mohseni, Zarei, and Ragan (2018):
  - Novice users
  - Domain experts
  - AI experts

Purpose of the paper

- Unify
  - Need for objective measure *and* subjective measure
  - More information might increase degree of explainability, but too much information elements result in information overlad
- Use of
  - Partial order relations
  - Measure theory
  - Law of diminishing returns (from economics)

Purpose of the paper

- Unify
  - Need for objective measure *and* subjective measure
  - More information might increase degree of explainability, but too much information elements result in information overlad
- Use of
  - Partial order relations
  - Measure theory
  - Law of diminishing returns (from economics)

Partial order relations

Illustration of partial order

- Example of partial order:

$$f \leq g \quad \Leftrightarrow \quad f(x) \leq g(x), \forall x \in [0,1]$$

- Application: $f(x) = x$ and $g(x) = 1 - x$
  - Neither $f \leq g$ nor $g \leq f$ holds

Explainability as partial order relation

- Suggestion: not all explanations are comparable
  - E.g. explanations meant for different categories of users
  - E.g. explanations that are logically non compatible

Partial order relations

Illustration of partial order

- Example of partial order:

$$f \leq g \quad \Leftrightarrow \quad f(x) \leq g(x), \forall x \in [0,1]$$

- Application: $f(x) = x$ and $g(x) = 1 - x$
  - Neither $f \leq g$ nor $g \leq f$ holds

Explainability as partial order relation

- Suggestion: not all explanations are comparable
  - E.g. explanations meant for different categories of users
  - E.g. explanations that are logically non compatible

Measure theory

Background on measure theory

- Measure theory is concerned with the "size" $m(E)$ of a set E
- Basic principle: $E_1 \subseteq E_2 \Rightarrow m(E_1) \leq m(E_2)$

Use of measure theory

- Degree of explainability as measure of explanation E?

$$E = I_1 \cup \ldots \cup I_n \quad \Rightarrow \quad m(E) = m(I_1 \cup \ldots \cup I_n)$$

- But:
  - Contradicts information overload and explanation selection
  - How to define measure of individual information elements?
    - Use of ontologies

Measure theory

Background on measure theory

- Measure theory is concerned with the "size" $m(E)$ of a set E
- Basic principle: $E_1 \subseteq E_2 \Rightarrow m(E_1) \leq m(E_2)$

Use of measure theory

- Degree of explainability as measure of explanation E?

$$E = I_1 \cup \ldots \cup I_n \quad \Rightarrow \quad m(E) = m(I_1 \cup \ldots \cup I_n)$$

- But:
  - Contradicts information overload and explanation selection
  - How to define measure of individual information elements?
    - Use of ontologies

Measure theory

Background on measure theory

- Measure theory is concerned with the "size" $m(E)$ of a set E
- Basic principle: $E_1 \subseteq E_2 \Rightarrow m(E_1) \leq m(E_2)$

Use of measure theory

- Degree of explainability as measure of explanation E?

$$E = I_1 \cup \ldots \cup I_n \quad \Rightarrow \quad m(E) = m(I_1 \cup \ldots \cup I_n)$$

- But:
  - Contradicts information overload and explanation selection
  - How to define measure of individual information elements?
    - Use of ontologies

Background on the law of diminishing returns

- Consider applying fertilizer to a corn field
- Will result in sharp increase in yield initially
- But from certain amount on, smaller increase in yield and eventually even decrease

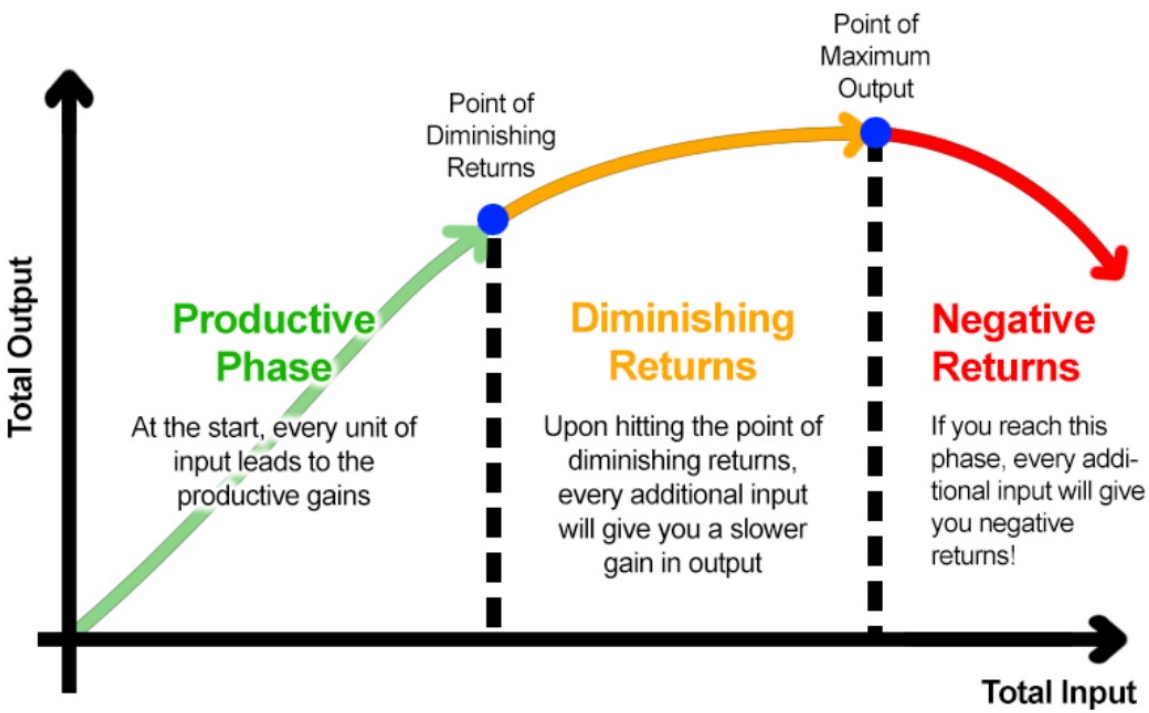

Figure: The law of diminishing returns

Use of the law of diminishing returns

- Consider sequence of information elements $I_1, \ldots, I_n$
- Consider $E_j = I_1 \cup \ldots \cup I_j$
- Consider $m(E_n)$ as an independent variable
  - Not as the degree of explainability
- Degree of explainability: function of the independent variable with curve agreeing with the law of diminishing returns

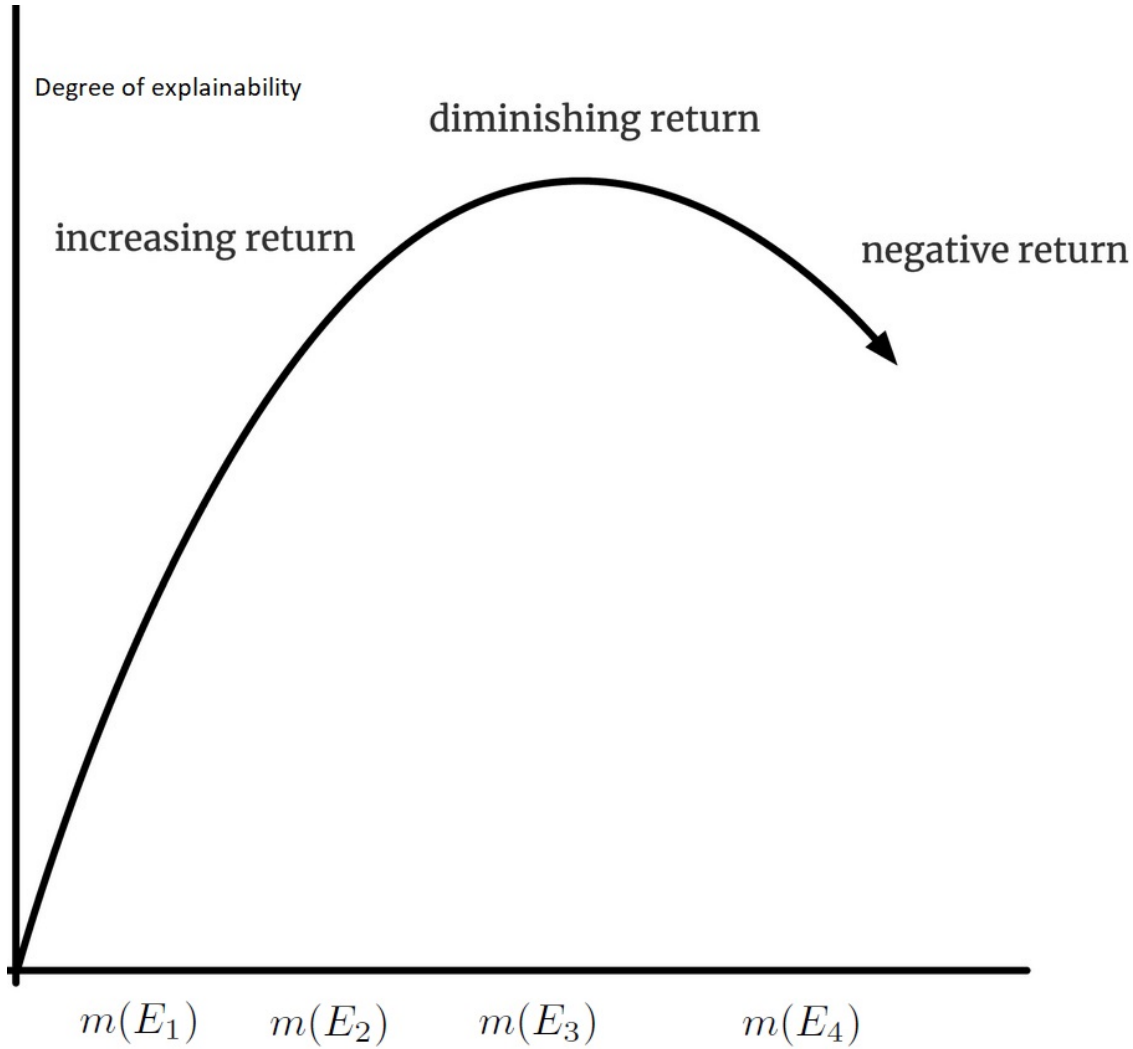

Figure: The law of diminishing returns in the context of explainability

- As more information elements are added, degree of explainability sharply increases, but then increase slows down, and eventually decrease
  - Takes into account information overload
  - Measure on X-axis and degree of explainability on Y-axis: unifies measure theory and explainability through law of diminishing returns
- Function is - in principle - numeric
  - Represents objective measure
- Function may be different for different categories of users
  - Represents subjective measure
  - Partial order relation

- As more information elements are added, degree of explainability sharply increases, but then increase slows down, and eventually decrease
  - Takes into account information overload
  - Measure on X-axis and degree of explainability on Y-axis: unifies measure theory and explainability through law of diminishing returns
- Function is - in principle - numeric
  - Represents objective measure
- Function may be different for different categories of users
  - Represents subjective measure
  - Partial order relation

- As more information elements are added, degree of explainability sharply increases, but then increase slows down, and eventually decrease
  - Takes into account information overload
  - Measure on X-axis and degree of explainability on Y-axis: unifies measure theory and explainability through law of diminishing returns
- Function is - in principle - numeric
  - Represents objective measure
- Function may be different for different categories of users
  - Represents subjective measure
  - Partial order relation

Conclusion

- Divergent aspects of explainability:
  - Objective *versus* subjective measures of explainability
  - More information means more explainable *versus* information overload
- These aspects can be unified through use of
  - Partial order relations
  - Measure theory
  - Law of diminishing returns

Future research

How to implement these ideas in practice?

Conclusion

- Divergent aspects of explainability:
  - Objective *versus* subjective measures of explainability
  - More information means more explainable *versus* information overload
- These aspects can be unified through use of
  - Partial order relations
  - Measure theory
  - Law of diminishing returns

Future research

How to implement these ideas in practice?

E-mail: **wim.demulder@ugent.be**