

Explainable AI and Bounded Rationality

Hersh Shefrin
Santa Clara University

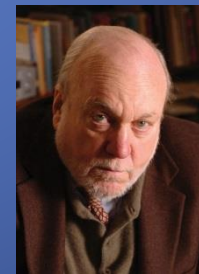
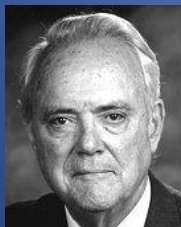
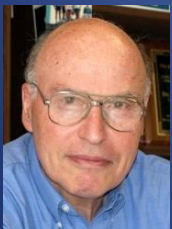
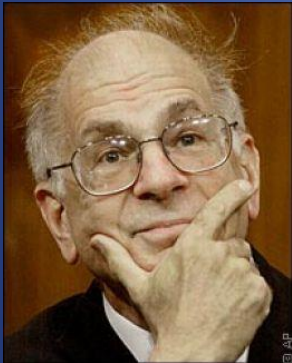
Herbert Simon Society
Turin, 2019



Outline

1. Herbert Simon.
2. AI revolution inflection point, Alpha Go.
3. U.S. Congressional hearings.
4. Explainable AI, compensatory structure.
5. Zest Finance, criticism of algorithm SHAP.
6. Insights from AI Fiddler.

1. Major Contributors to Psychology of Bounded Rationality



One Good Reason Decision Making

Ecological Rationality

What structure of information can Take The Best exploit?

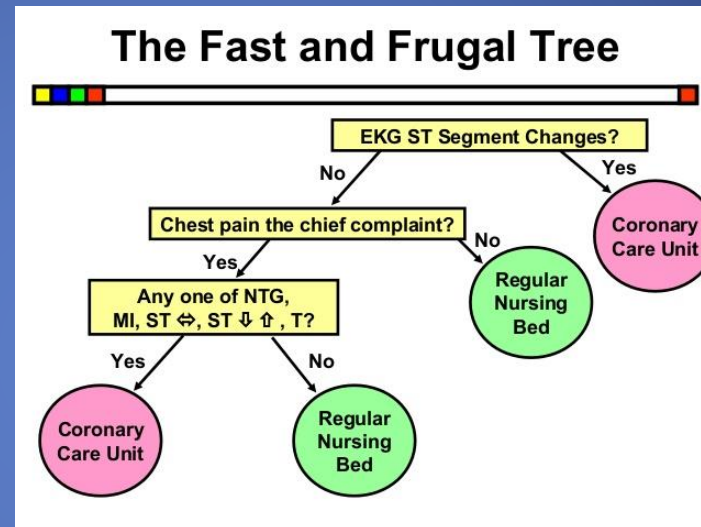


Result: If an environment consists of cues that are noncompensatory (e.g. $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and so forth), no weighted linear model can outperform the faster and more frugal *Take The Best*.

Martignon & Hoffrage (1999)

Fast & Frugal Trees

Green-Mehr fast & frugal improvement for admission to coronary care unit.



who should not be in the unit. Only 25 percent of the patients admitted to the coronary care unit did actually have a myocardial infarction (Green & Mehr, 1997; Green & Smith, 1988). Similar rates were found at larger hospitals (ranging from 12% to 42%).

Researchers at the University of Michigan Hospital tried to solve this overcrowding problem by training the physicians to use a decision-support tool based on logistic regression, rather than relying on their intuitive judgment (Green & Mehr, 1997).

2. Go & Heuristics

- In 2017, Google's AI algorithm AlphaGo won a three-match series against the world's best Go player Ke Jie.
- AlphaGo made its name in 2016 when it defeated high-profile Go player Lee Sedol 4-1.
- What does Alpha Go's success tell us about ecological rationality?

Bias? Who Won?

- Lee Sedol's comments and bias?
 - I'm **confident** about the match.
 - I believe that human intuition is still too advanced for AI to have caught up.




Smiling But Not Overconfident




3. U.S. Congressional Hearings 2019

financialservices.house.gov/calendar/eventsingle.aspx?EventID=403824







U.S. HOUSE COMMITTEE ON
FINANCIAL SERVICES



Chairwoman Maxine Waters

HOME ABOUT US NEWS COMMITTEE RESOURCES FORUMS CONSUMER HELP CONTACT



   

Hearings





Perspectives on Artificial Intelligence: Where We Are and the Next Frontier in Financial Services


Task Force on Artificial Intelligence

Subscribe for Updates

2128 RHOB,  Wednesday, June 26, 2019  10:00

Tags: [Task Force on Artificial Intelligence](#)

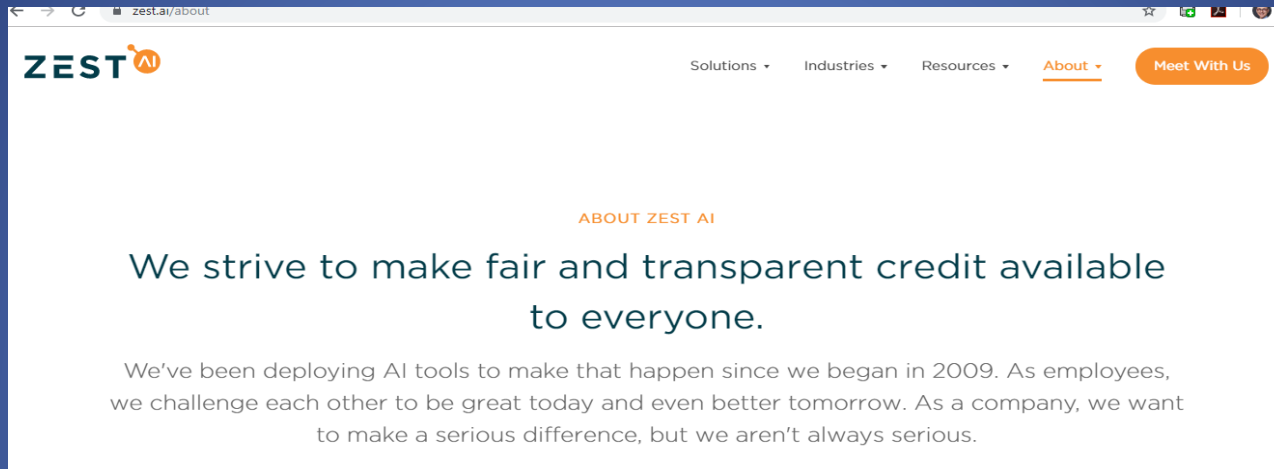
Twitter Feed  Follow Us

Douglas Merrill



My name is Douglas Merrill. I'm the CEO of ZestFinance, which I founded ten years ago with the mission to make fair and transparent credit available to everyone. Lenders use our software to increase loan approval rates, lower defaults, and make their lending fairer. Before ZestFinance, I was Chief Information Officer at Google. I have a Ph.D. in Artificial Intelligence from Princeton University.

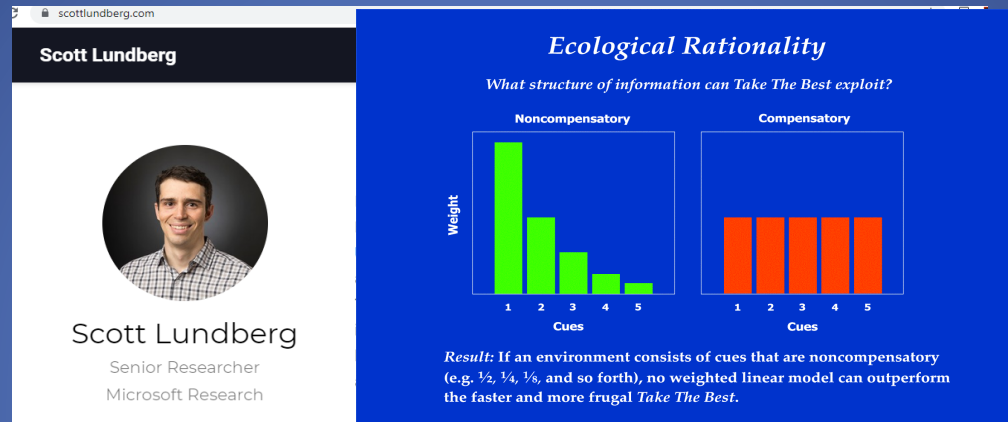
Fair Credit Reporting Act



- When lenders use AI to classify loan applicants into two subgroups, accept and reject, they face a legal issue.
- This is because when a consumer is denied credit, the Fair Credit Reporting Act of 1970 requires accurate and actionable reasons for the decision so that consumers can repair their credit and re-apply successfully.

4. Explainable AI

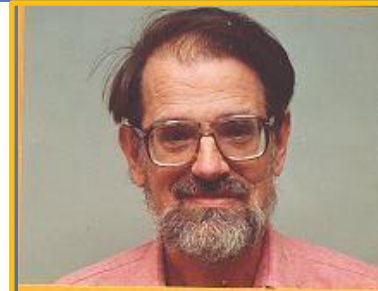
Shapley Value



The Shapley value takes as input a set function $v : 2^N \rightarrow R$. The Shapley value produces attributions s_i for each player $i \in N$ that add up to $v(N)$. The Shapley value of a player i is given by:

$$s_i = \sum_{S \subseteq N \setminus i} \frac{|S|! * (|N| - |S| - 1)!}{N!} (v(S \cup i) - v(S)) \quad (1)$$

There is an alternate permutation-based description of the Shapley value: Order the players uniformly at random, add them one at a time in this order, and assign to each player i its marginal contribution $V(S \cup i) - v(S)$; here S is the set of players that precede i in the ordering.



Theorem: $s()$ is the only function satisfying the four properties of Efficiency, Symmetry, Linearity and Null player.

Example, Cooperative Game

Example Shapley, players A & B			
<i>argument of $v()$</i>		$v()$	<i>marginal $v()$</i>
empty set		0	
A		1	1
A,B		1	0
empty set		0	
B		0	0
B,A		1	1
Shapley value B		0	
Shapley value A		1	

Application: treat cues as players and coalition value v as classification decision.

Example, Classification Task

- Next, consider a stylized example of a classification task, similar to the lending issues faced by firms such as Zest.
- This example will deal with recruitment/hiring.
- One of the features will be gender (m/f), to allow for issues of gender discrimination to arise.

Recruitment -- Hiring System

- A hiring manager interviews applicants for a job moving furniture.
- Applicants are rated on two major features, namely gender (m/f) and lifting ability (l/nl).
- The hiring manager uses a classification algorithm, that we can think of as an AI-black box.

Two Assistants

- The hiring manager has two assistants, one who exclusively monitors lifting ability and the other who exclusively monitors gender.
- The hiring manager tells us that he has hired 9 of every 10 applicants who have applied.

Information

- The assistant monitoring lifting ability tells us that the hiring manager has hired every applicant who qualifies as a lifter (l), and 4 out of every 5 non-lifters (nl).
- The assistant monitoring gender tells us that the hiring manager has hired every male applicant, and rejected all female applicants.

Explainability: Task for You

- For applicants who have been hired, if you had to divide 100 points between gender and lifting ability, to score the relative importance of these features, how would you divide the 100 points?
- This task is similar to how modern data scientists are approaching the question of explaining what AI black boxes are doing.

When AI Black Box Gives Us Fast & Frugal

- To provide additional intuition, suppose that the hiring manager's AI black box just happens to produce a fast & frugal heuristic.
- In the example to follow, the algorithm will be designated as f_{male} and as we shall see later satisfy the conditions for “take the best.”
- Of course, from the outside we do not know this, but want to use an explainability procedure that will inform us as much as possible about the nature of f_{male} .

Hiring System Games

SHAP

Two features/cues = gender (male) & strength (lift).

x_{male}	x_{lift}	$Pr [X = x]$	$f_{male}(x)$
0	0	10%	0
0	1	0%	0
1	0	40%	1
1	1	50%	1

Characteristic function for SHAP cooperative game

$v(\emptyset)$ = Expected prediction on the dataset D

$v(\{"male"\})$ = Conditional Expectation of the prediction when male=True - Expected prediction on the dataset D

$v(\{"lifter"\})$ = Conditional Expectation of the prediction when lifter=True - Expected prediction on the dataset D

$v(\{"male", "lifter"\})$ = 1 - Expected prediction on the dataset D

$v()$ in SHAP for f_{male}

$$v_{\mathbf{x}}^{\text{cond}}(S) = \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{\text{inp}}} [f(\mathbf{z}(\mathbf{x}, \mathbf{R}, S)) \mid \mathbf{R}_S = \mathbf{x}_S] - \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{\text{inp}}} [f(\mathbf{R})]$$

$$\mathbb{E}f(\mathbf{R}) = 0.9$$

= Prob of hiring taking place under f_{male} , b/c 90% of scenarios feature a male being available

S = male			
xmale = 1			
R takes on two values in above equation, namely (1,0) and (1,1)			
$f(1,0,S=\text{male})$	1		
$f(1,1,S=\text{male})$	1		
$\Pr(x_S = 1)$	0.9		
$\Pr(R_l = 0 \mid x_S = 1)$	44.4%		
$\Pr(R_l = 1 \mid x_S = 1)$	55.6%		
$\mathbb{E}(f(\mathbf{z}(1, R_l, S=\text{male})))$	1		
$v_{\mathbf{x}}^{\text{cond}}(S=\text{male})$	0.1		

Every male
hired (left);
4 of 5 non-
lifters hired
(right).

S = lift			
xlift=0			
$f(0,0,S=\text{lift})$	0		
$f(1,0,S=\text{lift})$	1		
$\Pr(x_l = 0)$	0.5		
$\Pr(R_m = 0 \mid x_S = 0)$	0.2		
$\Pr(R_m = 1 \mid x_S = 0)$	0.8		
$\mathbb{E}(f(\mathbf{z}(R_m, \text{lift}, S=\text{lift})))$	0.8		
$v_{\mathbf{x}}^{\text{cond}}(S=\text{lift})$	-0.1		

5. Zest Critique of SHAP

Jay Budzick, CTO Zest

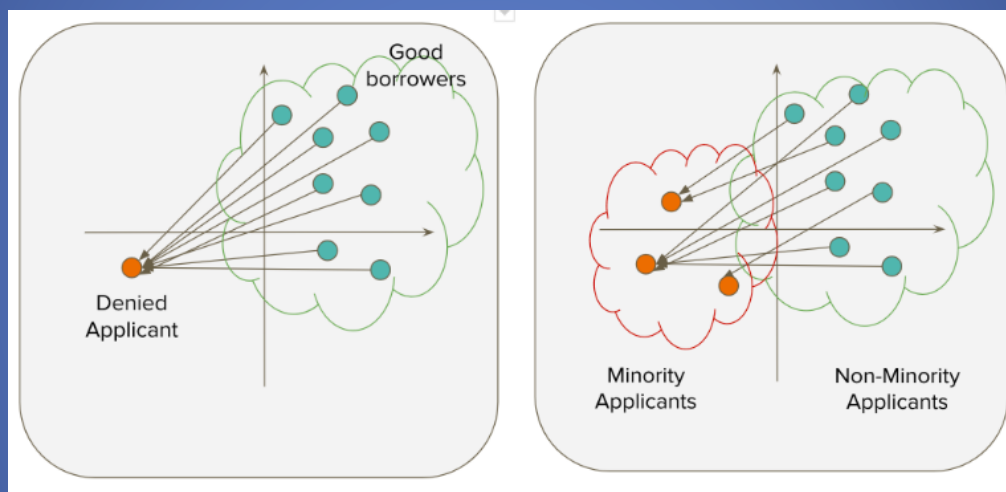
Why Lenders Shouldn't 'Just Use SHAP' To Explain Machine Learning Credit Models

Posted on May 30, 2019 by Jay Budzick



- There are many details you need to get right in this process, including the appropriate application of sample weights, mapping to score space at the approval cut-off, sampling methods, and accompanying documentation.
- Out of the box, SHAP doesn't allow you to easily do this.

- When lenders compute the reasons an applicant was rejected (for adverse action notices), they want to explain the applicant's score in terms of the approved applicants.



- Left: Adverse action requires comparing the denied applicant to good borrowers.
- Right: Fair lending analysis requires comparing minority applicants with non-minority applicants.

Scott Lundberg's Response

- Below are my thoughts on the recent Zest article "Why Not Just Use SHAP?".
- I think both Jay and John have a solid understanding of both finance and explainability and are doing good things.
- Unfortunately, I think this particular article was influenced too much by "marketing pressures" that encourage Zest to position itself as the go-to for explainable finance.

The many Shapley values for model explanation

Mukund Sundararajan (Google), Amir Najmi (Google)

ABSTRACT

The Shapley value has become a popular method to attribute the prediction of a machine-learning model on an input to its base features. The Shapley value [1] is known to be the unique method that satisfies certain desirable properties, and this motivates its use.

a service, and this incurs some cost. The Shapley value distributes this cost among the players. There is a correspondence between cost-sharing and the attribution problem: The cost function is analogous to the model, the players to base features, and the cost-shares to the attributions.

Algorithm 1 Computing CES on Training Data

Inputs are the Explicand x and the Examples T , each over a features set N

{Initializations}

for all $i \in N$ do
$$s_i \leftarrow 0 \text{ \{ } s_i \text{ is attribution for feature } i \text{ \}}$$

$T_i \leftarrow \{\}$ $\{T_i$ are examples that agree with explicand on feature $i\}$

end for

$$v_a \leftarrow 0 \text{ \{ } v_a \text{ is the average function value over } T \text{ \}}$$

```
{ // Take a pass over  $T$  to compute  $T_i$ 's and  $v_a$ }
```

for all $t \in T$ do
$$v_a \leftarrow v_a + f(x^t)/|T|$$
for all $i \in N$ do

if $x_i = x_i^t$ then

$$T_i = T_i \cup t$$

end if

end for

end for

```
{// Compute Shapley values via permutations}
```

for all permutations σ of N dofor all $i \in 1, \dots, |N|$ do

{The first element of the permutation is a special case.}

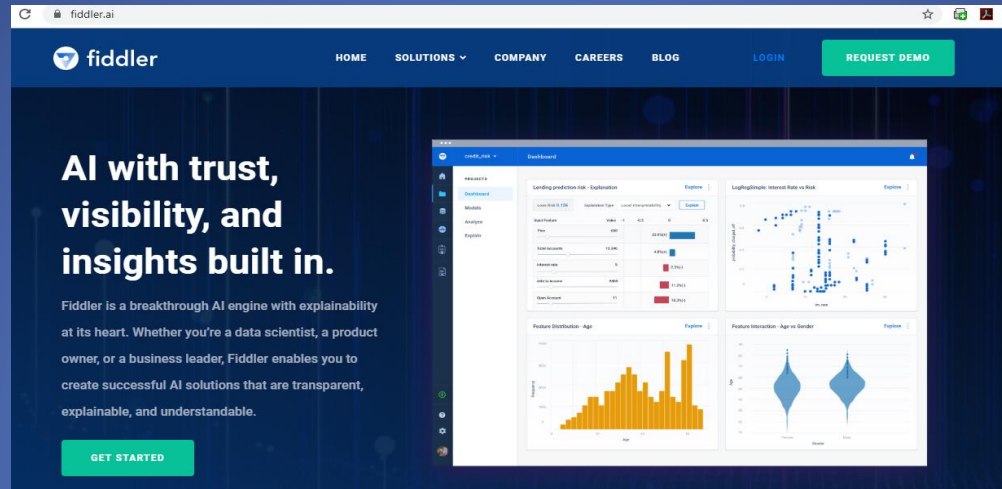
```

if i = 1 then

```

 $\{v_p = v(S), \text{ where } S \text{ consists of first } i - 1 \text{ features of } \sigma\}$

6. Insights from AI Fiddler



The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory

Abstract

Recently, a number of techniques have been proposed to explain a machine learning (ML) model's prediction by attributing it to the corresponding input features. Popular among these are techniques that apply the Shapley value method from cooperative game theory. While existing papers focus on the axiomatic motivation of Shapley values, and efficient

based on Shapley values from cooperative game theory being prominent among them.

Shapley values (Shapley 1953) provide a mathematically fair and unique method to attribute the payoff of a cooperative game to the players of the game. Due to its strong axiomatic guarantees, the Shapley values method is emerging as the de facto approach to feature attribution, and some

SHAP for f_{male}

$$v_{\mathbf{x}}^{\text{cond}}(S) = \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{\text{inp}}} [f(\mathbf{z}(\mathbf{x}, \mathbf{R}, S)) \mid \mathbf{R}_S = \mathbf{x}_S] - \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{\text{inp}}} [f(\mathbf{R})]$$

$$\mathbb{E}f(\mathbf{R}) = 0.9$$

= Prob of hiring taking place under f_{male} , b/c 90% of scenarios feature a male being available

S = male			
xmale = 1			
R takes on two values in above equation, namely (1,0) and (1,1)			
f(1,0,S=male)	1		
f(1,1,S=male)	1		
Pr(xS = 1)	0.9		
Pr(RI = 0 xS=1)	44.4%		
Pr(RI = 1 xS=1)	55.6%		
E(f(z(1,RI,S=male)))	1		
vxcond(S=male)	0.1		

xmale=1, xlift=1			
argument of v()	v	marginal v	
empty set	0.9		
add male	1	0.1	
add lift	1	0	
empty set	0.9		
add lift	1	0.1	
add male	1	0	
Shapley value lift	0.05		
Shapley value male	0.05		



Attribution Bias?

- f_{male} conforms to a “one good reason” decision rule.
- The “reason” / feature is gender, not strength.
- Yet, SHAP accords equal “value” to gender and strength.
- The unconditional probability of “hire” is 0.9.
- SHAP bias → Being male **only adds** 0.05 to the probability.
- SHAP bias → Being a lifter adds 0.05 to the probability.
- Being a “male lifter” adds $0.1 = 0.05 + 0.05$ to the base probability, with the total being 1.0.

Rejected Female Non-lifters

xmale=0, xlift=0		
argument of $v()$	v	marginal v
empty set	0.9	
add female	0	-0.9
add non-lift	0	0
empty set	0.9	
add non-lift	0.8	-0.1
add female	0	-0.8
Shapley value non-lift	-0.05	
Shapley value female	-0.85	

- The sum of -0.05 and -0.85 indicates the probability reduction from the 90% mean.
- Biased attribution: The attribution analysis indicates that some of the rejection stems from being a non-lifter.

Advice to Females

- What do you tell women applicants whose applications have been rejected?
- What do you tell them about what they can do to get hired under f_{male} ?
- What's the point here?!
- In this example, SHAP is unable to identify/**explain** the true nature of f_{male} .

Fast & Frugal Tree

- Consider a slight change to the example.
- Suppose that the hiring manager's AI algorithm is a fast & frugal tree designated as f_{both} .
- Again, neither the hiring manager nor we actually know the true character of f_{both} .
- The explainability task is to figure out what f_{both} is doing based on how we see it perform.

SHAP for f_{both}

- $f_{\text{both}} \sim$ a fast and frugal tree.
- SHAP \rightarrow For male lifters, most of the increase of 50%, from the average of 50%, stems from being a lifter.

$Ef(R) = 0.5$			
xmale=1, xlift=1			
argument of $v()$		v	marginal v
empty set		0.50	
add male		0.56	0.06
add lift		1.0	0.44
empty set		0.5	
add lift		1.0	0.5
add male		1.0	0
Shapley value lift		0.472	
Shapley value male		0.028	

Advice to Rejected Female Nonlifters

xmale=0, xlift=0			
argument of $v()$		v	marginal v
empty set		0.5	
add female		0	-0.5
add non-lift		0	0
empty set		0.5	
add non-lift		0	-0.5
add female		0	0
Shapley value non-lift		-0.25	
Shapley value female		-0.25	

- You can bring your odds of being hired from 0 to 25% by becoming a lifter.
- Still, $25\% <$ mean of 50%.

Female Lifters / Footballers?



SANTA CLARA UNIVERSITY
THE JESUIT UNIVERSITY IN SILICON VALLEY

[ABOUT SCU](#) | [ACADEMICS](#) | [ADMISSION](#) | [ATHLETICS](#) | [CAMPUS LIFE](#) | [GIVING](#)

Brandi Chastain '91 Making Headlines

[Home](#) › [News & Events](#) › [Feature Stories](#) › [2016 Feature Stories](#) › [Stories](#) › Brandi Chastain '91 Making Headlines



- Females can lift.
- What if they apply?
- Biased probabilities, b/c assigned prob is 0.

SHAP and f_{both} : Advice for Women

xmale=0, xlift=1		
argument of $v()$	v	marginal v
empty set	0.5	
add female	0	-0.5
add lift	0	0
empty set	0.5	
add lift	1	0.5
add female	0	-1
Shapley value lift	0.25	
Shapley value female	-0.75	

- As before, together they take the applicant's score down to 0 by their sum from a base of 0.5.
- As before, biased attribution to gender, but now too strong.

What's the Point?

- In this example, where neither the hiring manager nor we know the true nature of the AI black box algorithm, using SHAP to make AI explainable results in distortion, and less than full explainability!
- SHAP might just be too complex.
- If so, can a simpler approach do better?

SHAP is a Heuristic!

Less is More: Only Use Marginals

$$v_{\mathbf{x}}^{inp}(S) = \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{inp}} [f(\mathbf{z}(\mathbf{x}, \mathbf{R}, S))] - \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{inp}} [f(\mathbf{R})]$$

x_{male}	x_{lift}	$Pr[X = \mathbf{x}]$	$f_{male}(\mathbf{x})$
0	0	10%	0
0	1	0%	0
1	0	40%	1
1	1	50%	1

	x_{male}	x_{lift}
0	10%	50%
1	90%	50%

f_{male}

S = lift	
x _{lift} = 1	
f(0,1,S=lift)	0
f(1,1,S=lift)	1
Pr(x _l = 1)	0.5
Pr(R _m = 0 x _S =1)	0
Pr(R _m = 1 x _S =1)	1
E(f(z(R _m , lift, S=lift)))	1
vxcond(S=lift)	0.1

S = lift	
x _{lift} = 1	
f(0,1,S=lift)	0
f(1,1,S=lift)	1
Pr(x _l = 1)	0.5
Pr(R _m =0)	0.1
Pr(R _m =1)	0.9
E(f(z(R _m , lift, S=lift)))	0.9
vxcond(S=lift)	0.9

x _{male} =1, x _{lift} =1		
argument of v()	v	marginal v
empty set	0.9	
add male	1	0.1
add lift	1	0
empty set	0.9	
add lift	0.9	0
add male	1	0.1
Shap val lift	0.0	
Shap val male	0.1	

Less is More

Tally: Only Use Uniform Marginals

$$v_{\mathbf{x}}^{inp}(S) = \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{inp}} [f(z(\mathbf{x}, \mathbf{R}, S))] - \mathbb{E}_{\mathbf{R} \sim \mathcal{D}^{inp}} [f(\mathbf{R})]$$

<i>x</i> male	<i>x</i> lift	<i>Pr</i> [<i>X</i> = <i>x</i>]	<i>f</i> male(<i>x</i>)
0	0	10%	0
0	1	0%	0
1	0	40%	1
1	1	50%	1

	<i>x</i> male	<i>x</i> lift
0	50%	50%
1	50%	50%

fmale

S = lift	
<i>x</i> lift = 1	
<i>f</i> (0,1, <i>S</i> =lift)	0
<i>f</i> (1,1, <i>S</i> =lift)	1
<i>Pr</i> (<i>x</i> l = 1)	0.5
<i>Pr</i> (<i>R</i> m = 0 <i>x</i> <i>S</i> =1)	0
<i>Pr</i> (<i>R</i> m = 1 <i>x</i> <i>S</i> =1)	1
<i>E</i> (<i>f</i> (<i>z</i> (<i>R</i> m, lift, <i>S</i> =lift)))	1
<i>vx</i> cond(<i>S</i> =lift)	0.1

S = lift	
<i>x</i> lift = 1	
<i>f</i> (0,1, <i>S</i> =lift)	0
<i>f</i> (1,1, <i>S</i> =lift)	1
<i>Pr</i> (<i>x</i> l = 1)	0.5
<i>Pr</i> (<i>R</i> m = 0)	0.5
<i>Pr</i> (<i>R</i> m = 1)	0.5
<i>E</i> (<i>f</i> (<i>z</i> (<i>R</i> m, lift, <i>S</i> =lift)))	0.5
<i>vx</i> cond(<i>S</i> =lift)	0.5

<i>x</i> male=1, <i>x</i> lift=1		
<i>argument of v()</i>	<i>v</i>	<i>marginal v</i>
empty set	0.5	
add male	1	0.5
add lift	1	0
empty set	0.5	
add lift	0.5	0
add male	1	0.5
Shap val lift	0.0	
Shap val male	0.5	

Additional Thoughts About Bias

- The Explainable AI algorithms all use probabilistic information.
- The true Shapley value approach has a tree structure, with different x -values associated with base root branches.
- The true Shapley values are vectors, with sub-vectors corresponding to base root branches.
- Knowledge of probabilities ***even needed?***!

Previous Example, Cooperative Game

Example Shapley, players A & B			
<i>argument of $v()$</i>		$v()$	<i>marginal $v()$</i>
empty set		0	
A		1	1
A,B		1	0
empty set		0	
B		0	0
B,A		1	1
Shapley value B		0	
Shapley value A		1	

No probabilities needed!

Not in Explainable AI Literature

- All of the explainable AI algorithms use the full set of features as inputs at each stage, even if only to take expectations.
- In a game theoretic setting, a player i adds value by joining a coalition $S \setminus \{i\}$.
- The natural extension of this idea is for $v(S \setminus \{i\})$ to represent hiring managers providing hiring forecasts based only knowing about $S \setminus \{i\}$.

Joy Buolamwini *Frontline* Interview

On Biased AI Facial Recognition

- Even if you create some system you believe is somehow more objective, it's being used by humans at the end of the day.
- What I am saying is we also have to accept the fact that being human we're going to miss something.
- We're not going to get it all right.



Apple Card Managed by Goldman Sachs



DHH ✓ @dhh · Nov 7, 2019



The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.



Steve Wozniak ✓
@stevewoz

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

♡ 4,193 4:51 PM - Nov 9, 2019



What's the Point?

- This is like a real world version of f_{male} .
- Gender discrimination is clear.
- What is the Apple Card algorithm doing?
- That's the explainable AI question.
- Better for providers like Goldman-Sachs to be doing decent explainability analysis before going to market.

Summary

1. Herbert Simon.
2. AI revolution inflection point, Alpha Go.
3. U.S. Congressional hearings.
4. Explainable AI, compensatory structure.
5. Zest Finance, criticism of SHAP.
6. Insights from AI Fiddler.