

4.2(എ) അൽഗോരിതങ്ങൾ പഠിക്കുന്നതിൽ നിന്ന് ധാർമ്മികമായി എതിർക്കപ്പെടേണ്ട പക്ഷപാതങ്ങൾ എങ്ങനെ തടയാം?

ചരിത്രത്തിൽ ഭൂരിഭാഗവും, മറ്റുള്ളവരെ ബാധിക്കുന്ന കാര്യങ്ങളെക്കുറിച്ച് ആളുകൾ തീരുമാനങ്ങൾ എടുത്തിട്ടുണ്ട്. നിയമനം, വായ്പ എടുക്കൽ, ആളുകൾക്ക് എന്ത് രോഗങ്ങളുണ്ടെന്ന് കണ്ടെത്തൽ, ചിലറ്റു വിൽപന, നിർമ്മാണം, വിനോദം തുടങ്ങിയ കാര്യങ്ങളെക്കുറിച്ച് ആളുകൾ തീരുമാനങ്ങൾ എടുത്തു. എന്നാൽ കഴിഞ്ഞ കുറച്ച് ദശകങ്ങളിൽ, ആർട്ടിഫിഷ്യൽ ഇന്റലിജൻസ് (AI) ആളുകളെക്കാൾ മികച്ചതും വിശ്വസനീയവുമായ ചില കാര്യങ്ങൾ ചെയ്യാൻ കഴിഞ്ഞു. ഉദാഹരണത്തിന്, 1997-ൽ, IBM നിർമ്മിച്ച കമ്പ്യൂട്ടർ ചെസ്സ് പ്രോഗ്രാമായ ഡീപ്പ് ബ്ലൂ ലോകത്തിലെ ഏറ്റവും മികച്ച ചെസ്സ് കളിക്കാരനായ ഗാരി കാസ്പറോവിനെ പരാജയപ്പെടുത്തി. നിയമനം, വായ്പ യോഗ്യത, ഭവനം, വൈദ്യശാസ്ത്രം, സാങ്കേതികവിദ്യ, ബാങ്കിംഗ്, മാർക്കറ്റിംഗ്, വിനോദം തുടങ്ങിയ മേഖലകളിൽ തീരുമാനങ്ങൾ എടുക്കാൻ ഇപ്പോൾ AI ഉപയോഗിക്കുന്നു. ആളുകൾക്ക് കഴിയുന്നതിനേക്കാൾ കൃത്യമായി ചില ജോലികൾ ചെയ്യാൻ AI-ക്ക് കഴിയുമെന്നതിനാൽ ഈ മേഖലകളിൽ AI ഉപയോഗിക്കുന്നു. നിയമനം, വിനോദം, മറ്റ് വ്യവസായങ്ങൾ എന്നിവയിൽ AI അൽഗോരിതങ്ങൾക്ക് കാര്യക്ഷമത വർദ്ധിപ്പിക്കാൻ കഴിയുമെങ്കിലും, ഈ അൽഗോരിതങ്ങൾ പക്ഷപാതത്തിനും വിവേചനത്തിനും കാരണമാകും, പ്രത്യേകിച്ചും പക്ഷപാതപരമായ ഡാറ്റയിൽ പരിശീലനം നേടുന്നോഴോ അവ നടപ്പിലാക്കുന്നതിൽ ശരിയായ മേൽനോട്ടം ഇല്ലാതിരിക്കുന്നോഴോ, ഇത് ചില ഗ്രൂപ്പുകൾക്ക് അന്യായമായ ഫലങ്ങൾക്ക് കാരണമാകും. ഡാറ്റയിൽ നിന്ന് പാറ്റേണുകൾ പഠിക്കാനും തീരുമാനങ്ങളോ പ്രവചനങ്ങളോ എടുക്കാനും ഒരു കമ്പ്യൂട്ടർ പിന്തുടരുന്ന ഒരു ഘട്ടം ഘട്ടമായുള്ള നിയമങ്ങളുടെ ഒരു കൂട്ടമാണ് ഒരു അൽഗോരിതം. AI സിസ്റ്റങ്ങളിൽ, സിസ്റ്റം എന്താണ് പഠിക്കുന്നത്, എങ്ങനെ പഠിക്കുന്നു, എങ്ങനെ പ്രവർത്തിക്കുന്നു എന്നിവ അൽഗോരിതങ്ങൾ നിർണ്ണയിക്കുന്നു.

ചരിത്ര രേഖകൾ മുൻകാല മുൻവിധികൾ പ്രകടിപ്പിച്ചെക്കാം, ഉപയോക്താ പെരുമാറ്റം സാമൂഹിക മുൻധാരണകളെ പ്രതിഫലിപ്പിച്ചെക്കാം, കൂടാതെ മനുഷ്യൻ വ്യാഖ്യാനിച്ച വിവരങ്ങൾ ആത്മനിഷ്ഠമായ വിലയിരുത്തലുകളോ അവ്യക്തമായ പക്ഷപാതങ്ങളോ ഉൾക്കൊള്ളിച്ചെക്കാം. പരിശീലന ഡാറ്റ സാമൂഹികമോ സാംസ്കാരികമോ സ്ഥാപനപരമോ ആയ പക്ഷപാതങ്ങൾ പ്രകടിപ്പിക്കുമ്പോൾ പക്ഷപാതം "സ്വഭാവത്തിന്റെയോ മനോഭാവത്തിന്റെയോ ഒരു മുൻതൂക്കം" ആയി നിർവ്വചിക്കപ്പെടുന്നു. തൽഫലമായി, ഒരു സ്ഥാപനത്തെ മറ്റൊന്നിനേക്കാൾ മുൻഗണനയോടെ പരിഗണിക്കുന്നത് പക്ഷപാതമായി മാറുന്നു. വാസ്തവത്തിൽ, പക്ഷപാതം വിവേചനവുമായി അന്തർലീനമായി ബന്ധപ്പെട്ടിരിക്കുന്നു, ഇത് "മുൻവിധിയോടെയുള്ളതോ മുൻവിധിയോടെയുള്ളതോ ആയ വീക്ഷണം, പെരുമാറ്റം അല്ലെങ്കിൽ ചികിത്സ" ആയി വിശേഷിപ്പിക്കപ്പെടുന്നു.

"പക്ഷപാതത്തിന്റെ അനന്തരഫലമാണ് വിവേചനം; വികലമായ ഒരു വീക്ഷണം മറ്റുള്ളവരോടുള്ള വിവേചനപരമായ പെരുമാറ്റത്തിന് കാരണമാകും.

ചരിത്രപരമായ രേഖകൾ, ഉപയോക്താ പെരുമാറ്റം, മനുഷ്യൻ വ്യാഖ്യാനിച്ച ഡാറ്റാസെറ്റുകൾ എന്നിവയുൾപ്പെടെയുള്ള ഉറവിടങ്ങളിൽ നിന്നാണ് AI സിസ്റ്റങ്ങൾ അറിവ് നേടുന്നത്. അൽഗോരിതങ്ങൾ പാറ്റേണുകൾ തിരിച്ചറിയുന്ന രീതിയെയും പ്രവചനങ്ങൾ സൃഷ്ടിക്കുന്ന രീതിയെയും തിരഞ്ഞെടുപ്പുകൾ നൽകുന്ന രീതിയെയും ഈ ഡാറ്റാ ഉറവിടങ്ങൾ സ്വാധീനിക്കുന്നു, അടിസ്ഥാന ഡാറ്റ സാമൂഹിക അസമത്വങ്ങളെ പ്രതിഫലിപ്പിക്കുന്നുവെങ്കിൽ ഇത് പക്ഷപാതപരമായ ഫലങ്ങളിലേക്ക് നയിച്ചെക്കാം. സാമൂഹിക, സാംസ്കാരിക, സ്ഥാപന സാഹചര്യങ്ങളിൽ സൃഷ്ടിക്കപ്പെടുന്ന ഡാറ്റ, നിലവിലുള്ള മാനുഷിക മൂല്യങ്ങളെയും വിശ്വാസങ്ങളെയും അനീതികളെയും പലപ്പോഴും പ്രതിഫലിപ്പിക്കുന്നു; അതിനാൽ, AI സിസ്റ്റങ്ങൾ ഈ പക്ഷപാതങ്ങളെ പുനർനിർമ്മിക്കുകയും വർദ്ധിപ്പിക്കുകയും ചെയ്യുന്നു, പകരം

അവയെ തിരുത്തുന്നു. നിലവിലുള്ള ഡാറ്റയിൽ നിന്ന് പാറ്റേണുകൾ AI സ്വാംശീകരിക്കുമ്പോൾ, ലിംഗ പക്ഷപാതം, വംശീയ മുൻവിധി അല്ലെങ്കിൽ സാമ്പത്തിക അസമത്വം പോലുള്ള സാമൂഹിക വിവേചനപരമായ രീതികൾ അൽഗോരിതമിക് തീരുമാനമെടുക്കലിൽ വേരൂന്നിയിരിക്കുന്നു. തൽഫലമായി, റിക്രൂട്ട്മെന്റ്, നിയമ നിർവ്വഹണം, ക്രെഡിറ്റ് അസസ്മെന്റ്, ആരോഗ്യ സംരക്ഷണം തുടങ്ങിയ മേഖലകളിൽ AI അസമമായ ഫലങ്ങൾ സൃഷ്ടിച്ചെക്കാം, തുല്യത വളർത്തുന്നതിനുപകരം നിലവിലുള്ള അസമത്വങ്ങൾ നിലനിർത്തുന്നു.

കാര്യകാരണ ബന്ധവുമായി പരസ്പരബന്ധം സ്ഥാപിക്കുന്നതിലൂടെയോ, ആത്മനിഷ്ഠമായ വ്യക്തിഗത സ്വഭാവസവിശേഷതകളുടെ സ്വാധീനത്തിലൂടെയോ, പക്ഷപാതപരമോ അപര്യാപ്തമോ ആയ ഡാറ്റാ സെറ്റുകളെയോ മനുഷ്യ സ്രോതസ്സുകളെയോ ആശ്രയിക്കുന്നതിലൂടെയോ ഈ പക്ഷപാതം ഉണ്ടാകാം. AI പക്ഷപാതം പരിഹരിക്കാനുള്ള ശ്രമം തുടർച്ചയായി നടക്കുന്നു, ഈ മേഖലയിൽ നേരായ പരിഹാരങ്ങളൊന്നുമില്ല. നിർദ്ദിഷ്ട ജനസംഖ്യാശാസ്ത്രത്തിനെതിരായ വിവേചനം സൂക്ഷ്മമായി അവതരിപ്പിക്കപ്പെടാൻ സാധ്യതയുള്ളതിനാൽ AI മുൻവിധി നിയമപരമായ വെല്ലുവിളികൾ നേരിടാൻ സാധ്യതയുണ്ട്. വൈവിധ്യമാർന്ന തീരുമാനമെടുക്കൽ പ്രക്രിയകളിൽ സമൂഹം കൂടുതലായി AI ഉപയോഗിക്കുന്നതിനാൽ, പക്ഷപാതത്തെക്കുറിച്ചുള്ള ആശങ്ക അവഗണിക്കരുത്.

ആർട്ടിഫിഷ്യൽ ഇന്റലിജൻസ് (AI) സംവിധാനങ്ങൾ ഡാറ്റയിൽ നിന്ന് പാറ്റേണുകൾ വേർതിരിച്ചറിയുന്നു. എന്നിരുന്നാലും, ഡാറ്റ നിലവിലുള്ള സാമൂഹിക അസമത്വങ്ങളോ പക്ഷപാതങ്ങളോ ഉൾക്കൊള്ളുന്നുവെങ്കിൽ, അൽഗോരിതം അവയെ സ്വാംശീകരിക്കുകയും വഷളാക്കുകയും ചെയ്തേക്കാം, ഇത് വംശീയ പ്രൊഫൈലിംഗ്, ലിംഗപരമായ സ്റ്റീറിയോടൈപ്പിംഗ് അല്ലെങ്കിൽ അസമമായ തൊഴിൽ രീതികൾ പോലുള്ള ധർമ്മികമായി പ്രശ്നകരമായ മുൻവിധികളിലേക്ക് നയിച്ചേക്കാം. വംശീയ പ്രൊഫൈലിംഗ് എന്നത് പ്രത്യേക വസ്തുതകളെ അടിസ്ഥാനമാക്കിയല്ല, പ്രധാനമായും അവരുടെ വംശം, വംശീയത അല്ലെങ്കിൽ വംശീയ സ്വഭാവവിശേഷങ്ങൾ എന്നിവയെ അടിസ്ഥാനമാക്കിയുള്ള കുറ്റകൃത്യങ്ങളുടെയോ പരിശോധനയുടെയോ സംശയത്തിന് വ്യക്തികളെ ഒറ്റപ്പെടുത്തുന്ന മുൻവിധിപരമായ രീതിയെ സൂചിപ്പിക്കുന്നു. തൽഫലമായി, അത്തരം പക്ഷപാതങ്ങൾ ലഘൂകരിക്കുന്നത് ഒരു സാങ്കേതികവും ധർമ്മികവുമായ പ്രശ്നമാണ്.

വിവേചനത്തിന്റെ ഒരു മാതൃകാപരമായ ഉദാഹരണമാണ് വംശീയത, ഒരു പ്രത്യേക വംശത്തിലെ ഒരു വ്യക്തി തന്റെ വംശത്തിലെ അംഗങ്ങളുമായുള്ള ഇടപെടലുകളെ അപേക്ഷിച്ച് മറ്റ് വംശങ്ങളിലെ അംഗങ്ങളുമായി വ്യത്യസ്തമായി ഇടപഴകുന്നു. തൽഫലമായി, വ്യത്യസ്ത വംശത്തിലെ ഒരു വ്യക്തിയെ കണ്ടുമുട്ടിയാൽ, അവർ അവജ്ഞയോടെയും ശത്രുതയോടെയും ആക്രമണോത്സുകതയോടെയും പോലും പ്രതികരിക്കുന്നു. വംശീയത പ്രവർത്തനത്തിൽ പ്രകടിപ്പിക്കപ്പെടാത്ത ഒരു മനോഭാവമായി നിലനിൽക്കുമോ എന്ന് കാണാൻ ഇത് നിർബന്ധിതമായ ഒരു അന്വേഷണമാണ്. വംശീയത അനിവാര്യമായും ഒരു പ്രവൃത്തിയിൽ കലാശിക്കുന്നില്ലെന്ന് സൂചിപ്പിക്കുന്ന ഈ ചോദ്യത്തിന് സ്ഥിരീകരണമായി ഉത്തരം നൽകാൻ കഴിയുമെന്ന് ഞാൻ വാദിക്കുന്നു. അതിനാൽ, സ്വഭാവപരമായ വംശീയത - സ്വന്തം വംശത്തിലെ വ്യക്തികളെ പരിഗണിക്കുന്നതിനേക്കാൾ വ്യത്യസ്തമായി പരിഗണിക്കാനുള്ള പ്രവണത - പ്രായോഗികമായി വംശീയത എന്നിവ തമ്മിൽ നമുക്ക് വേർതിരിച്ചറിയാൻ കഴിയും.

തീരുമാനമെടുക്കൽ അൽഗോരിതം ബാഹ്യ ഘടകങ്ങൾ പരിഗണിക്കുമ്പോൾ, പക്ഷപാതം നിർദ്ദിഷ്ട ഗ്രൂപ്പുകൾക്കെതിരെ വിവേചനം സൃഷ്ടിക്കുന്നു. പ്രസക്തമായ പാരാമീറ്ററുകൾ ഉൾപ്പെടുത്തുന്നത് കൃത്യത വർദ്ധിപ്പിക്കും, അതേസമയം അപ്രസക്തമായ പാരാമീറ്ററുകൾ കൃത്യതയെ ദോഷകരമായി ബാധിക്കുകയും "വംശീയത, ലിംഗവിവേചനം, മറ്റ് അനീതികൾ" എന്നിവ വർദ്ധിപ്പിക്കുകയും ചെയ്യും, ഇത് പാർശ്വവൽക്കരിക്കപ്പെട്ട വിഭാഗങ്ങൾക്ക് അന്യായമായി ദോഷം വരുത്തുന്ന വളച്ചൊടിച്ച് ഫലങ്ങളിലേക്ക് നയിക്കുന്നു. അപ്രസക്തമായ ഘടകങ്ങൾ ആത്യന്തിക ലക്ഷ്യത്തിൽ സ്വാധീനം ചെലുത്താത്തവയെ ഉൾക്കൊള്ളുന്നു. ദേശീയത, വംശം പോലുള്ള വ്യക്തിഗത ഘടകങ്ങൾ ഒരു വ്യക്തിയുടെ കഴിവുകളെ സ്വാധീനിക്കുന്നില്ല. ഒരു അൽഗോരിതത്തിന് ഈ അറിവ് നൽകുന്നത് അതിന്റെ കൃത്യത വർദ്ധിപ്പിക്കില്ല, കൂടാതെ മുൻവിധി അല്ലെങ്കിൽ

വിവേചനം ഈ ഘടകങ്ങളെ AI തീരുമാനമെടുക്കലിന് ഹാനികരമാക്കുന്നു, കാരണം അവ നിലവിലുള്ള അസമത്വങ്ങൾ നിലനിർത്തുന്നതിലൂടെ ന്യൂനപക്ഷങ്ങൾ അല്ലെങ്കിൽ പ്രാതിനിധ്യം കുറഞ്ഞ സമൂഹങ്ങൾ പോലുള്ള ചില വിഭാഗങ്ങളെ അന്യായമായി പ്രതികൂലമായി ബാധിക്കുന്ന പക്ഷപാതപരമായ ഫലങ്ങളിലേക്ക് നയിച്ചേക്കാം.

ഒരു ഗ്രൂപ്പിന്റെ പൊതുവായ സാംസ്കാരിക, ഭാഷാ, പൂർവ്വിക, അല്ലെങ്കിൽ സാമൂഹിക സവിശേഷതകളെയാണ് വംശീയത സൂചിപ്പിക്കുന്നത്. അൽഗോരിതങ്ങൾ അന്യായമായി പ്രതികൂലമായി ബാധിക്കുമ്പോഴോ പ്രത്യേക വംശീയ വിഭാഗങ്ങളിൽ നിന്നുള്ള വ്യക്തികളെ തെറ്റായി ചിത്രീകരിക്കുമ്പോഴോ ആണ് AI സിസ്റ്റങ്ങളിൽ വംശീയ പക്ഷപാതം ഉണ്ടാകുന്നത്.

AI-യിൽ വംശീയതയുമായി ബന്ധപ്പെട്ട പക്ഷപാതം എങ്ങനെ ഉണ്ടാകുന്നു:

പക്ഷപാതപരമായ പരിശീലന ഡാറ്റ: ചരിത്രപരമായ ഡാറ്റ ചില വംശീയ വിഭാഗങ്ങളുടെ വിവേചനമോ പ്രാതിനിധ്യക്കുറവോ പ്രതിഫലിപ്പിച്ചേക്കാം.

സാമ്പിൾ ബയസ്: ഡാറ്റാബേസുകൾ (മുഖ തിരിച്ചറിയൽ ഡാറ്റാസെറ്റുകൾ പോലുള്ളവ) ചില വംശീയ വിഭാഗങ്ങളെ വേണ്ടത്ര പ്രതിനിധീകരിക്കണമെന്നില്ല.

ഡിസൈൻ തിരഞ്ഞെടുപ്പുകൾ: സവിശേഷതകളോ ലേബലുകളോ വംശീയ മാർക്കറുകളെ പരോക്ഷമായി എൻകോഡ് ചെയ്തേക്കാം, ഇത് അസമമായ ഫലങ്ങളിലേക്ക് നയിച്ചേക്കാം.

ആഘാതങ്ങൾ:

ചില വംശീയ വിഭാഗങ്ങൾക്ക് മുഖം തിരിച്ചറിയിൽ കൃത്യമല്ല.

നിയമനം, ക്രെഡിറ്റ് സ്കോറിംഗ്, പോലീസിംഗ് അല്ലെങ്കിൽ ആരോഗ്യ സംരക്ഷണം എന്നിവയിലെ വിവേചനപരമായ ഫലങ്ങൾ വിശ്വാസ്യത നഷ്ടപ്പെടുകയും സാമൂഹിക അസമത്വങ്ങൾ ശക്തിപ്പെടുത്തുകയും ചെയ്യുന്നു.

അസന്തുലിതമായ പരിശീലന ഡാറ്റാ കാരണം, പ്രത്യേക വംശീയ ഉത്ഭവത്തിൽ നിന്നുള്ള വ്യക്തികളിൽ മുഖം തിരിച്ചറിയിൽ അൽഗോരിതങ്ങൾ ഉയർന്ന പിശക് നിരക്കുകൾ കാണിക്കുന്നു, ഇത് തുല്യതയെയും പൗരസ്വാതന്ത്ര്യത്തെയും കുറിച്ചുള്ള ആശങ്കകൾ ഉയർത്തുന്നു. തീരുമാനമെടുക്കുന്നതിനുള്ള ഒരു AI മോഡൽ വ്യത്യസ്ത ഗ്രൂപ്പുകൾക്കിടയിൽ അബദ്ധവശാൽ ഉണ്ടാകുന്ന പരസ്പരബന്ധങ്ങളെ തിരിച്ചറിഞ്ഞേക്കാം, അതുവഴി നിയമനം അല്ലെങ്കിൽ നിയമപാലനം പോലുള്ള മേഖലകളിൽ അന്യായമായ പെരുമാറ്റത്തിലേക്ക് നയിച്ചേക്കാവുന്ന വംശീയ അല്ലെങ്കിൽ ലിംഗ പക്ഷപാതം പോലുള്ള ഈ ഗ്രൂപ്പുകൾക്കെതിരായ വിവേചനം തീവ്രമാക്കും. തൽഫലമായി, AI മോഡലുകളിൽ നിന്ന് വ്യക്തിഗത സവിശേഷതകൾ ഇല്ലാതാക്കുന്നത് നിർദ്ദിഷ്ട ജനസംഖ്യയ്ക്കെതിരായ വിവേചനം ലഘൂകരിക്കാൻ സഹായിക്കുന്നു. തൽഫലമായി, AI മോഡലുകളിൽ നിന്ന് വ്യക്തിഗത സവിശേഷതകൾ ഇല്ലാതാക്കുന്നത് നിർദ്ദിഷ്ട ജനസംഖ്യയ്ക്കെതിരായ വിവേചനം ലഘൂകരിക്കാൻ സഹായിക്കുന്നു. ഒരു അൽഗോരിതത്തിന്റെ ലക്ഷ്യ ജനസംഖ്യയെ ഉചിതമായി പ്രതിഫലിപ്പിക്കുന്നതിൽ പരാജയപ്പെടുന്ന വികലമായ ഡാറ്റാ സെറ്റുകളിൽ നിന്നാണ് ഓട്ടോമേറ്റഡ് തീരുമാനമെടുക്കലിൽ പക്ഷപാതത്തിന്റെ ഒരു പ്രസക്തമായ ഉറവിടം ഉണ്ടാകുന്നത്. തൽഫലമായി, ഡാറ്റാ സെറ്റിനുള്ളിൽ പ്രാതിനിധ്യമില്ലാത്ത ജനസംഖ്യാശാസ്ത്രവുമായി ബന്ധപ്പെട്ട പിശകുകളുടെ ഉയർന്ന സംഭവങ്ങൾ അൽഗോരിതങ്ങൾ പ്രകടിപ്പിച്ചേക്കാം, ഇത് അന്യായമായ പെരുമാറ്റത്തിനും ആ ഗ്രൂപ്പുകൾക്ക് ദോഷകരമായ ഫലങ്ങൾക്കും കാരണമാകാം, ഉദാഹരണത്തിന് പ്രവചന പോലീസിംഗിലോ പക്ഷപാതപരമായ വായു അംഗീകാര പ്രക്രിയകളിലോ തെറ്റായ നെഗറ്റീവുകളുടെ വർദ്ധിച്ച നിരക്കുകൾ. അൽഗോരിതങ്ങൾ വികസിപ്പിക്കുന്നതിൽ ഉൾപ്പെട്ടിരിക്കുന്ന മനുഷ്യരുടെ മുൻവിധികളിൽ നിന്നാണ് AI പക്ഷപാതങ്ങൾ ഉത്ഭവിച്ചത്. തൊഴിൽ വിപണിയിലെ പക്ഷപാതങ്ങൾക്ക് മനുഷ്യർ സംഭാവന നൽകുന്നു. "ആഫ്രിക്കൻ അമേരിക്കൻ പേരുകൾ" ഉള്ളവരെ അപേക്ഷിച്ച് "വെള്ളക്കാരായ പേരുകൾ" ഉൾപ്പെടുന്ന റെസ്യൂമുകൾക്ക് 50 ശതമാനം കൂടുതൽ അഭിമുഖ കോൾബാക്കുകൾ ലഭിക്കുന്നുവെന്ന് ഒരു പഠനം വെളിപ്പെടുത്തി (ബെർട്രാൻഡ് & മുള്ളെനാഥൻ, 2003, പേജ് 2). ഈ പരീക്ഷണം തെളിയിച്ചു തൊഴിൽ വിപണിയിലെ മനുഷ്യ വിവേചനം. കൃത്രിമബുദ്ധി, കമ്പ്യൂട്ടർ അൽഗോരിതങ്ങൾ, ഡാറ്റാ ശേഖരണം എന്നിവയിലൊക്കെ മനുഷ്യ ഘടകങ്ങൾ ഉൾപ്പെട്ടിരിക്കുന്നതിനാൽ, കമ്പ്യൂട്ടർ, AI സിസ്റ്റങ്ങളിൽ മനുഷ്യന്റെ മുൻവിധികൾ എങ്ങനെ പ്രതിഫലിക്കാമെന്ന് വ്യക്തമാണ്, ഇത് വിവേചനത്തിലേക്ക് നയിക്കുന്നു, ഇത് പക്ഷപാതപരമായ നിയമന രീതികൾ, അസമമായ വേതനം, ചില ഗ്രൂപ്പുകൾക്ക് പരിമിതമായ അവസരങ്ങൾ എന്നിവയിൽ പ്രകടമാകാം.

ഒരു AI സിസ്റ്റത്തിലെ പക്ഷപാതത്തിന്റെ ഒരു ഉദാഹരണം Amazon.com, Inc. ന്റെ 2014 ലെ AI നിയമന സംരംഭത്തിൽ നിന്നാണ് ഉത്ഭവിച്ചത്. തൊഴിൽ അൽഗോരിതത്തിൽ സ്ത്രീകൾക്കെതിരായ മുൻവിധികൾ തിരിച്ചറിഞ്ഞതിനെത്തുടർന്ന് ആമസോൺ ഈ സംരംഭം അവസാനിപ്പിക്കാൻ നിർബന്ധിതരായി. ആമസോണിന് സമർപ്പിച്ച ഒരു ദശാബ്ദക്കാലത്തെ റെസ്യൂമുകൾ ഉൾപ്പെടുന്ന പരിശീലന ഡാറ്റയിൽ നിന്നാണ് വിവേചനം ഉണ്ടായത്. എന്നിരുന്നാലും, സാങ്കേതികവിദ്യയിലെ ലിംഗ നിയമന അസമത്വത്തിന്റെ ഫലമായി "ടെക് വ്യവസായത്തിനുള്ളിൽ പുരുഷ ആധിപത്യം" (ഖണ്ഡിക 6) നിലനിൽക്കുന്നതിനാൽ, അൽഗോരിതം പുരുഷന്മാരിൽ നിന്നുള്ള റെസ്യൂമെകളോടും "സ്ത്രീകൾ" എന്ന പദം ഉൾക്കൊള്ളുന്ന ശിക്ഷിക്കപ്പെട്ട റെസ്യൂമെകളോടും ഒരു പക്ഷപാതം വികസിപ്പിച്ചെടുത്തു.

മുൻപറഞ്ഞ ഉദാഹരണം അതിന്റെ പക്ഷപാതപരവും മോശം തൊഴിൽ അൽഗോരിതവും കൊണ്ട് മാത്രമല്ല, ദീർഘകാലമായി നിലനിൽക്കുന്ന ലിംഗ വിവേചനം എന്ന പ്രശ്നം കൂടുതൽ വഷളാക്കുന്നതിനാലും ശ്രദ്ധേയമാണ്. AI തീരുമാനമെടുക്കൽ സാധ്യതയുള്ള ജീവനക്കാരെ എങ്ങനെ പ്രതികൂലമായി ബാധിക്കുമെന്നും വിശാലമായ സാമൂഹിക വിവേചന പ്രശ്നങ്ങൾ എങ്ങനെ വഷളാക്കുമെന്നും ഈ കേസ് വ്യക്തമാക്കുന്നു.

ലിംഗഭേദം എന്നത് വ്യക്തികളോടുള്ള അന്യായമായ അല്ലെങ്കിൽ അസമമായ പെരുമാറ്റത്തെ സൂചിപ്പിക്കുന്നു, ഇത് അവരുടെ

ലിംഗഭേദത്തെ അടിസ്ഥാനമാക്കിയുള്ള അന്യായമായ അല്ലെങ്കിൽ അസമമായ പെരുമാറ്റത്തെ സൂചിപ്പിക്കുന്നു. ഇത് വിവിധ ദൈനംദിന സാഹചര്യങ്ങളിൽ ഉയർന്നുവരുന്നു. അവസരങ്ങളെയും അവകാശങ്ങളെയും സാമൂഹിക ഇടപെടലുകളെയും ബാധിക്കുന്നു. സ്ത്രീകൾ ശാസ്ത്ര-സാങ്കേതിക പഠനങ്ങളിൽ ഏർപ്പെടുന്നതിൽ നിന്ന് നിരുത്സാഹപ്പെടുത്തിയേക്കാം, അതേസമയം പുരുഷന്മാർ വികാരങ്ങൾ പ്രകടിപ്പിക്കുന്നതിൽ നിന്നോ പരിചരണ തൊഴിലുകൾ തിരഞ്ഞെടുക്കുന്നതിൽ നിന്നോ തടയപ്പെട്ടേക്കാം. താരതമ്യപ്പെടുത്താവുന്ന തൊഴിലിന് പുരുഷന്മാരേക്കാൾ താഴ്ന്ന പ്രതിഫലം സ്ത്രീകൾക്ക് ലഭിച്ചേക്കാം, സ്ഥാനക്കയറ്റത്തിന് തടസ്സങ്ങൾ നേരിടാം, അല്ലെങ്കിൽ ജോലി അഭിമുഖങ്ങളിൽ വൈവാഹിക, ശിശു സംരക്ഷണ ഉത്തരവാദിത്തങ്ങളെക്കുറിച്ചുള്ള അന്വേഷണങ്ങൾ നേരിടാം. മിക്ക ആളുകളും സ്ത്രീകൾ ഗാർഹിക ജോലിയും പരിചരണവും ഏറ്റെടുക്കുമെന്ന് പ്രതീക്ഷിക്കുന്നു, അതേസമയം പുരുഷന്മാർ ഗാർഹിക ഉത്തരവാദിത്തങ്ങളിൽ തുല്യമായി പങ്കെടുക്കുന്നതിൽ നിന്ന് അവർ നിരുത്സാഹപ്പെടുത്തിയേക്കാം. സ്ത്രീകൾക്ക് പീഡനമോ ചലനശേഷിയിൽ നിയന്ത്രണങ്ങളോ നേരിടാം, അതേസമയം പുരുഷന്മാർക്ക് ശക്തിയുടെയും വൈകാരിക അടിച്ചമർത്തലിന്റെയും ആദർശങ്ങളുമായി പൊരുത്തപ്പെടാനുള്ള സമ്മർദ്ദം അനുഭവപ്പെടാം, ഇത് പരിചരണത്തിലും വിടുജോലികളിലും തുല്യമായി ഏർപ്പെടാനുള്ള അവരുടെ കഴിവിനെ തടസ്സപ്പെടുത്തും. പുരുഷന്മാരെ നേതാക്കളായും സ്ത്രീകളെ പരിചരണകരായും ചിത്രീകരിക്കുന്നത് ലിംഗപരമായ സ്റ്റീരിയോടൈപ്പുകളെയും സാമൂഹിക പ്രതീക്ഷകളെയും ശക്തിപ്പെടുത്തുന്നു.

AI സിസ്റ്റങ്ങൾ വ്യക്തികളെ അവരുടെ ലിംഗഭേദം അടിസ്ഥാനമാക്കി അന്യായമായി വിലയിരുത്തുമ്പോഴാണ് AI-യിൽ ലിംഗ വിവേചനം സംഭവിക്കുന്നത്. പക്ഷപാതപരമായ ഡാറ്റ, ഡിസൈൻ തിരഞ്ഞെടുപ്പുകൾ അല്ലെങ്കിൽ സിസ്റ്റത്തിനുള്ളിൽ ഉൾച്ചേർത്ത സാമൂഹിക പക്ഷപാതങ്ങൾ എന്നിവ കാരണം അൽഗോരിതങ്ങൾ പുരുഷന്മാർക്കും സ്ത്രീകൾക്കും അല്ലെങ്കിൽ നോൺ-ബൈനറി വ്യക്തികൾക്കും അസമമായ ഫലങ്ങൾ സൃഷ്ടിക്കുമ്പോഴാണ് ഇത് സംഭവിക്കുന്നത്, ഇത് നിയമനം, വായ്പ നൽകൽ, നിയമ നിർവ്വഹണം തുടങ്ങിയ മേഖലകളിൽ കാര്യമായ ദോഷങ്ങൾക്ക് കാരണമാകാം.

സാങ്കേതിക മേഖലകളിലും നേതൃത്വ സ്ഥാനങ്ങളിലും പുരുഷ മേധാവിത്വം ഉൾപ്പെടെയുള്ള അസമമായ സാമൂഹിക റോളുകൾ പ്രകടമാക്കുന്ന ചരിത്രപരമായ ഡാറ്റയിൽ നിന്നാണ് ലിംഗപരമായ മുൻവിധി പലപ്പോഴും ഉണ്ടാകുന്നത്. AI അത്തരം ഡാറ്റ പ്രോസസ്സ് ചെയ്യുമ്പോൾ, അത് സ്റ്റീരിയോടൈപ്പുകളെ ശക്തിപ്പെടുത്തും - പുരുഷന്മാരെ ഉയർന്ന ശമ്പളമുള്ള തൊഴിലുകളിലും സ്ത്രീകളെ പരിചരണ റോളുകളിലും ബന്ധിപ്പിക്കുന്നു. അൽഗോരിതങ്ങളുടെ രൂപീകരണവും വികസന ടീമുകളിലെ ലിംഗ വൈവിധ്യത്തിന്റെ അഭാവവും ഈ പക്ഷപാതത്തെ തീവ്രമാക്കും, ഇത് റിക്രൂട്ട്മെന്റ്, ക്രെഡിറ്റ്, ആരോഗ്യ സംരക്ഷണം, പരസ്യം ചെയ്യൽ എന്നിവയിൽ വിവേചനപരമായ ഫലങ്ങളിലേക്ക് നയിക്കും, ഇത് നിലവിലുള്ള അസമത്വങ്ങൾ നിലനിർത്തുകയും പ്രാതിനിധ്യം കുറഞ്ഞ ഗ്രൂപ്പുകൾക്ക് അവസരങ്ങൾ പരിമിതപ്പെടുത്തുകയും ചെയ്യും.

ഈ തരത്തിലുള്ള മുൻവിധി AI-യിൽ തന്നെ ആരോപിക്കപ്പെടുന്നില്ല, കാരണം അത് മനുഷ്യ സ്രോതസ്സുകളിൽ നിന്നാണ് ഉണ്ടാകുന്നത്. ചുരുക്കത്തിൽ, കൃത്രിമബുദ്ധിയെ ട്രെയിനിംഗിൽ നിന്ന് ഒഴിവാക്കുന്നത് മനുഷ്യ പ്രേരിത പക്ഷപാതത്തെ അന്തർലീനമായി ഇല്ലാതാക്കുകയോ കുറയ്ക്കുകയോ ചെയ്യില്ല. ഉപസംഹാരമായി, പരസ്പരബന്ധം, ബാഹ്യമായ വ്യക്തിഗത ഘടകങ്ങൾ, നിർദ്ദിഷ്ട ജനസംഖ്യാശാസ്ത്രത്തെ ഒഴിവാക്കുന്ന വളഞ്ഞതോ അപൂർണ്ണമോ ആയ ഡാറ്റ സെറ്റുകൾ, അല്ലെങ്കിൽ അൽഗോരിതം വികസിപ്പിച്ച വ്യക്തികളുടെ പക്ഷപാതങ്ങൾ എന്നിവയെ അടിസ്ഥാനമാക്കിയുള്ള തെറ്റായ കാര്യകാരണ അനുമാനങ്ങളിൽ നിന്ന് ഒരു പക്ഷപാതപരമായ അൽഗോരിതം ഉരുവിച്ചേക്കാം. ഈ സവിശേഷതകളെല്ലാം പക്ഷപാതത്തിന് കാരണമായേക്കാം, ഇത് നിലവിലുള്ള വിവേചനം തീവ്രമാക്കും.

AI തീരുമാനങ്ങൾ സാമൂഹിക മുൻവിധി നിലനിർത്തുന്ന കാര്യമായ പക്ഷപാതങ്ങൾ പ്രകടിപ്പിച്ചേക്കാമെന്ന് ഞങ്ങൾ കണ്ടെത്തി. കാരണ-കാരണബന്ധം പരസ്പരബന്ധവുമായി കൂട്ടിക്കുഴയ്ക്കുക, ഒരു അൽഗോരിതത്തിന്റെ തീരുമാനമെടുക്കുമായി ബന്ധമില്ലാത്ത ഘടകങ്ങളെ ആശ്രയിക്കുക, പക്ഷപാതപരമായ ഡാറ്റ സെറ്റുകൾ ഉപയോഗിക്കുക, അല്ലെങ്കിൽ മനുഷ്യ ഇൻപുട്ട് ഉൾപ്പെടുത്തുക എന്നിവയിൽ നിന്ന് അത്തരം പക്ഷപാതങ്ങൾ ഉണ്ടാകാം, ഇവയെല്ലാം നിയമനം, വായ്പ നൽകൽ, നിയമ നിർവ്വഹണം തുടങ്ങിയ വിവിധ ആപ്ലിക്കേഷനുകളിൽ അന്യായമായ ഫലങ്ങൾക്ക് കാരണമാകാം. ഓട്ടോമേറ്റഡ് തീരുമാനത്തിലെ പക്ഷപാതങ്ങൾ തിരുത്താൻ ഏതൊക്കെ രീതികൾ ഉപയോഗിക്കാം?

സംവിധാനങ്ങൾ ഉണ്ടാക്കുന്നുണ്ടോ?

പക്ഷപാതം ഇല്ലാതാക്കുക.

AI സിസ്റ്റങ്ങളിൽ നിന്നുള്ള മുൻവിധി ഇല്ലാതാക്കാൻ, മുകളിൽ പറഞ്ഞ എല്ലാ ബുദ്ധിമുട്ടുകളും പരിഹരിക്കേണ്ടത് അത്യാവശ്യമാണ്. ആദ്യത്തെ രണ്ട് ആശങ്കകൾ സങ്കീർണ്ണമായി ബന്ധപ്പെട്ടിരിക്കുന്നു: തിരഞ്ഞെടുത്ത വേരിയബിളുകൾക്ക് വിവേചന ശേഷിയുണ്ടെങ്കിൽ രണ്ട് പരസ്പരബന്ധിതമായ വേരിയബിളുകളിൽ നിന്നുള്ള കാര്യകാരണബന്ധം അനുമാനിക്കുന്നത് പക്ഷപാതത്തിന് കാരണമാകുന്നു. അൽഗോരിതത്തിന്റെ തീരുമാനമെടുക്കൽ പ്രക്രിയയെ സ്വാധീനിക്കാൻ പാടില്ലാത്ത "വ്യക്തിഗത ഘടകങ്ങൾ" ഈ വേരിയബിളുകളാണ്. എന്നിരുന്നാലും, അത്തരം വിവേചന സ്വഭാവസവിശേഷതകൾ നീക്കം ചെയ്യുന്നത് തോന്നിയേക്കാവുന്നതിനേക്കാൾ സങ്കീർണ്ണമാണ്.

ഒരു അൽഗോരിതത്തിലേക്ക് നൽകുന്ന ഡാറ്റാ പക്ഷപാതപരമായ മാനുഷിക വിധിന്യായങ്ങളെ ഉൾക്കൊള്ളുന്നതോ ചരിത്രപരമോ സാമൂഹികമോ ആയ അസമത്വങ്ങളെ പ്രതിഫലിപ്പിക്കുന്നതോ ആകാം, ലിംഗഭേദം, വംശീയത, ലൈംഗിക ആഭിമുഖ്യം തുടങ്ങിയ സെൻസിറ്റീവ് സ്വഭാവസവിശേഷതകളുടെ അഭാവത്തിൽ പോലും (മാന്വീക തുടങ്ങിയവർ, 2019, ഖണ്ഡിക 4). ഉദാഹരണത്തിന്, ആമസോണിന്റെ ലിംഗ-പക്ഷപാതപരമായ നിയമന അൽഗോരിതത്തിന്റെ കാര്യത്തിൽ, "നടപ്പിലാക്കി", "പിടിച്ചു" തുടങ്ങിയ പദങ്ങൾ സ്ത്രീകളോട് വിവേചനം കാണിക്കാൻ ഉപയോഗിച്ചു. കാരണം പുരുഷ സ്ഥാനാർത്ഥികളുടെ റെസ്യൂമുകളിൽ പലപ്പോഴും ഈ പദപ്രയോഗങ്ങൾ വലിയ അളവിൽ ഉൾപ്പെടുത്തിയിട്ടുണ്ട് (ഡാസ്റ്റിൻ, 2018). തൽഫലമായി, വ്യക്തമായ വ്യക്തിഗത വേരിയബിളുകൾ ഒഴികെ.

വിവേചനം ഒഴിവാക്കാൻ അൽഗോരിതങ്ങളിൽ നിന്നുള്ള വിവരങ്ങൾ പര്യാപ്തമല്ല. പരിശീലന ഡാറ്റയുടെ വിവിധ വശങ്ങളിൽ ഈ സവിശേഷതകൾ പ്രകടമായേക്കാം.

വികലമായ ഡാറ്റാ സെറ്റുകളിൽ നിന്ന് മറ്റൊരു പക്ഷപാത സ്രോതസ്സ് ഉയർന്നുവരുന്നു. യഥാർത്ഥ പക്ഷപാതങ്ങൾ മൂലമോ നിർദ്ദിഷ്ട ജനസംഖ്യയെ അപര്യാപ്തമായി പ്രതിനിധീകരിക്കുന്ന സ്ഥിതിവിവരക്കണക്കുകൾ മൂലമോ സ്റ്റാറ്റിസ്റ്റിക്കൽ സെറ്റുകൾ വളച്ചൊടിപ്പെട്ടേക്കാം. ഉദാഹരണത്തിന്, നിയമന രീതികളിലെ നിലവിലുള്ള ലിംഗ അസമത്വങ്ങൾ ഡാറ്റാസെറ്റിനെ വളച്ചൊടിച്ചതിനാൽ ആമസോണിന്റെ അൽഗോരിതം പക്ഷപാതം കാണിച്ചു (ഡാസ്റ്റിൻ, 2018), ഇത് തുല്യ യോഗ്യതയുള്ള സ്ത്രീ സ്ഥാനാർത്ഥികളേക്കാൾ പുരുഷ സ്ഥാനാർത്ഥികളെ അനുകൂലിക്കുന്നതിലേക്ക് നയിച്ചു.

ഈ സാഹചര്യത്തിൽ മുൻവിധി ലഘൂകരിക്കുന്നതിന്, മുൻ നിയമന പാറ്റേണുകളിൽ നിന്ന് വ്യതിചലിക്കുന്ന തീരുമാനങ്ങൾ AI പ്രോഗ്രാം നൽകണം. എല്ലാ ജനസംഖ്യാ ഗ്രൂപ്പുകളുടെയും അപര്യാപ്തമായ പ്രാതിനിധ്യം കാരണം ഡാറ്റാ സെറ്റുകൾ പക്ഷപാതം കാണിക്കുമ്പോൾ, അവ ശരിയാക്കണം. ഈ വിഷയത്തിന് വ്യത്യസ്ത കാഴ്ചപ്പാടുകളുണ്ട്. ഗുഗിളിന്റെ AI വകുപ്പ് പറയുന്നതനുസരിച്ച്, "വ്യക്തികളുടെ യഥാർത്ഥ ലോക ആവൃത്തികളെ കൂടുതൽ കൃത്യമായി പ്രതിനിധീകരിക്കുന്നതിന് പൊതു പരിശീലന ഡാറ്റാ സെറ്റുകൾക്ക് പലപ്പോഴും വർദ്ധനവ് ആവശ്യമായി വരും" ("ഉത്തരവാദിത്തമുള്ള AI പ്രാക്ടീസുകൾ", n.d., ഖണ്ഡിക 9). ഡാറ്റാ സെറ്റുകൾക്ക് അന്തർലീനമായി പക്ഷപാതം ഉണ്ടായിരിക്കാമെന്നും അത്തരം പക്ഷപാതം ലഘൂകരിക്കുന്നതിന് പരിഷ്കരണവും മെച്ചപ്പെടുത്തലും ആവശ്യമാണെന്നും ഈ വീക്ഷണം എടുത്തുകാണിക്കുന്നു, പ്രാതിനിധ്യം കുറഞ്ഞ ഗ്രൂപ്പുകളെ ഓവർസാമ്പിൾ ചെയ്യുകയോ കൂടുതൽ സന്തുലിതമായ പ്രാതിനിധ്യം ഉറപ്പാക്കാൻ സിന്ററ്റിക് ഡാറ്റാ ഉൾപ്പെടുത്തുകയോ പോലുള്ള സാങ്കേതിക വിദ്യകൾ ഇതിൽ ഉൾപ്പെടാം.

അൽഗോരിതങ്ങൾ പഠിച്ചുകൊണ്ട് ധാർമ്മികമായി അഭികാമ്യമല്ലാത്ത പക്ഷപാതങ്ങൾ ഉണ്ടാകുന്നത് ലഘൂകരിക്കുന്നതിന്, പ്രശ്ന വിവരണം, ഡാറ്റാ ശേഖരണം, മോഡൽ പരിശീലനം, വിന്യാസം എന്നിവയുൾപ്പെടെ മുഴുവൻ AI വികസന പ്രക്രിയയും ഉൾക്കൊള്ളുന്ന ഒരു സമഗ്രവും മൾട്ടി-ഫേസ് തന്ത്രവും ആവശ്യമാണ്. പക്ഷപാതം സാധാരണയായി ഒരു സാങ്കേതിക പിഴവല്ല, മറിച്ച് മനുഷ്യ ഡാറ്റയിൽ വേരുന്നിയ ചരിത്രപരമോ സാമൂഹികമോ ആയ അസമത്വങ്ങളുടെ പ്രകടനമാണെന്ന് തിരിച്ചറിയുക എന്നതാണ് അടിസ്ഥാന തത്വം.

ടാക്കിംഗ്

തൽഫലമായി, നീതിയും ഉത്തരവാദിത്തവും ഉറപ്പാക്കുന്നതിന്, ധാർമ്മികമായ AI വികസനത്തിന് സൂക്ഷ്മമായ ഡാറ്റാ തിരഞ്ഞെടുപ്പ്, പക്ഷപാത വിലയിരുത്തൽ, തുടർച്ചയായ മേൽനോട്ടം എന്നിവ ആവശ്യമാണ്.

ഡാറ്റാസെറ്റുകൾ ഓഡിറ്റ് ചെയ്യുന്നു

പ്രാതിനിധ്യത്തിനും നീതിക്കും വേണ്ടിയുള്ള ഡാറ്റാസെറ്റുകൾ ഓഡിറ്റ് ചെയ്യുന്നത്, വിവിധ സാമൂഹിക ഗ്രൂപ്പുകളെ ഉചിതമായും, ആനുപാതികമായും, തുല്യമായും പ്രതിനിധീകരിക്കുന്നുണ്ടെന്ന് ഉറപ്പാക്കാൻ പരിശീലന ഡാറ്റ സൂക്ഷ്മമായി പരിശോധിക്കുകയും വിലയിരുത്തുകയും ചെയ്യുന്നതിൽ ഉൾപ്പെടുന്നു. AI സിസ്റ്റങ്ങൾ വിവേചനപരമോ അന്യായമോ ആയ ഫലങ്ങൾ സൃഷ്ടിക്കാൻ കാരണമായേക്കാവുന്ന അസന്തുലിതാവസ്ഥ, ഒഴിവാക്കലുകൾ അല്ലെങ്കിൽ പക്ഷപാതപരമായ പ്രവണതകൾ കണ്ടെത്തുന്നതിന് ഈ സമീപനം സഹായിക്കുന്നു. മുൻവിധി കറയ്ക്കുന്നതിനും AI സിസ്റ്റങ്ങൾ ധർമ്മികമായും, സമഗ്രമായും, സാമൂഹികമായി ഉത്തരവാദിത്തത്തോടെയും പ്രവർത്തിക്കുന്നുവെന്ന് ഉറപ്പാക്കുന്നതിനും പതിവ് ഡാറ്റാസെറ്റ് ഓഡിറ്റുകൾ നിർണായകമാണ്.

ന്യായബോധം ഉൾച്ചേർക്കൽ

നീതിയും സമത്വവും കണക്കിലെടുത്ത് തീരുമാനമെടുക്കൽ പ്രക്രിയകളിലെ കൃത്യതയും കാര്യക്ഷമതയും കണക്കിലെടുത്ത് മനഃപൂർവ്വം അൽഗോരിതങ്ങൾ രൂപകൽപ്പന ചെയ്യുന്നതാണ് ന്യായമായ നിയന്ത്രണങ്ങൾ. വിവിധ സാമൂഹിക ഗ്രൂപ്പുകൾക്ക് (ഉദാ. ലിംഗഭേദം, ജാതി, വംശം) തുല്യമായ അവസരങ്ങളും ഫലങ്ങളും വാശാനം ചെയ്യാൻ ഈ സിസ്റ്റം ശ്രമിക്കുന്നു. പരമ്പരാഗത അൽഗോരിതങ്ങൾ കൃത്യത, കാര്യക്ഷമത അല്ലെങ്കിൽ ലാഭക്ഷമത എന്നിവയ്ക്ക് മാത്രമേ മുൻഗണന നൽകൂ. തുല്യത ഒരു ലക്ഷ്യമാകുമ്പോൾ, അൽഗോരിതങ്ങൾ ഫലപ്രാപ്തിയെ ന്യായവുമായി പൊരുത്തപ്പെടുത്തണം, അങ്ങനെ ഫലങ്ങൾ ഏതെങ്കിലും പ്രത്യേക ഗ്രൂപ്പിനോ വ്യക്തിക്കോ ആനുപാതികമല്ലാത്ത രീതിയിൽ ദോഷം വരുത്തുന്നില്ലെന്ന് ഉറപ്പാക്കണം. സമത്വം, മാനുഷ്യത, വിവേചനമില്ലായ്മ തുടങ്ങിയ ധർമ്മിക ആദർശങ്ങൾ അളവനുസരിച്ച് പ്രവർത്തിക്കുന്നു.

അൽഗോരിതത്തിനുള്ളിലെ മാനദണ്ഡങ്ങൾ.

നൈതിക ഒപ്റ്റിമൈസേഷൻ

സാങ്കേതിക പ്രകടനം (കൃത്യത, വേഗത, അല്ലെങ്കിൽ കാര്യക്ഷമത ഉൾപ്പെടെ) വർദ്ധിപ്പിക്കുക മാത്രമല്ല, ധർമ്മികവും സാമൂഹികവുമായ മൂല്യങ്ങളുമായി പൊരുത്തപ്പെടുന്ന അൽഗോരിതങ്ങളുടെ വികസനമാണ് നൈതിക ഒപ്റ്റിമൈസേഷൻ അർത്ഥമാക്കുന്നത്. നൈതിക മൂല്യങ്ങളെ പരിമിതികളോ സമകാലിക ലക്ഷ്യങ്ങളോ ആയി കണക്കാക്കുന്നു. സമത്വം, വിവേചനമില്ലായ്മ, സുതാര്യത, മനുഷ്യ അന്തസ്സ് തുടങ്ങിയ തത്വങ്ങൾ ഒപ്റ്റിമൈസേഷൻ പ്രക്രിയയിൽ സംയോജിപ്പിച്ചിരിക്കുന്നു. നൈതിക ഒപ്റ്റിമൈസേഷൻ സാമൂഹികവും സാംസ്കാരികവുമായ വശങ്ങൾ പരിഗണിക്കുന്നു, മനുഷ്യർ അവലോകനം ചെയ്യുകയും സാധൂകരിക്കുകയും ആവശ്യമുള്ളിടത്ത് അൽഗോരിതമിക് വിധിന്യായങ്ങളെ മറികടക്കുകയും ചെയ്യുന്നു. AI ഫലങ്ങളോടുള്ള മനുഷ്യന്റെ ഉത്തരവാദിത്തം അസന്ദിഗ്ധമായി സംരക്ഷിക്കപ്പെടുന്നു, അതിനാൽ കമ്പ്യൂട്ടറുകളിലേക്ക് ധർമ്മിക ഉത്തരവാദിത്തം കൈമാറുന്നത് തടയുന്നു. നിയമ നിർവ്വഹണം, നിയമനം, ആരോഗ്യ സംരക്ഷണം, ക്ഷേമ വിഹിതം തുടങ്ങിയ നിർണായക മേഖലകളിൽ നിർണായകമാണ്.

വിവേചനപരമായ സവിശേഷതകൾ പരിമിതപ്പെടുത്തൽ

വിവേചനപരമായ സ്വഭാവസവിശേഷതകളെ നിയന്ത്രിക്കുന്നതിലൂടെ, അൽഗോരിതങ്ങൾ സെൻസിറ്റീവ് സ്വഭാവസവിശേഷതകളോ അവയുടെ മറഞ്ഞിരിക്കുന്ന ബദലുകളോ ഉപയോഗിക്കുന്നതിൽ നിന്ന് വിലക്കുന്നു, ഇത് അന്യായമോ മുൻവിധിയോ ആയ ഫലങ്ങൾക്ക് കാരണമായേക്കാം. വംശം, ജാതി, മതം, ലിംഗഭേദം, വൈകല്യം തുടങ്ങിയ സ്വഭാവവിശേഷങ്ങൾ മോഡൽ പരിശീലന സമയത്ത് ഒഴിവാക്കുകയോ കർശനമായി നിയന്ത്രിക്കുകയോ ചെയ്യുന്നു. വംശം, ജാതി, മതം, ലിംഗഭേദം അല്ലെങ്കിൽ വൈകല്യം പോലുള്ള അന്തർലീനമോ സാമൂഹികമായി ആരോപിക്കപ്പെടുന്നതോ ആയ ഗുണങ്ങൾ കാരണം വ്യക്തികളോ ഗ്രൂപ്പുകളോ അന്യായമോ അസമമോ ആയ പെരുമാറ്റത്തിന് വിധേയമാകുമ്പോഴാണ് വിവേചനം സംഭവിക്കുന്നത്. അൽഗോരിതങ്ങളുടെയും കൃത്രിമബുദ്ധിയുടെയും മണ്ഡലത്തിൽ, അത്തരം മുൻവിധി പലപ്പോഴും വ്യവസ്ഥാപിതവും മനഃപൂർവ്വവുമാണ്, നിയമന പ്രക്രിയകളിൽ പക്ഷപാതപരമായ തീരുമാനമെടുക്കൽ, നിയമ നിർവ്വഹണം, സേവനങ്ങളിലേക്കുള്ള പ്രവേശനം എന്നിവയുൾപ്പെടെ ഗുരുതരമായ പ്രത്യാഘാതങ്ങൾ ഉണ്ടാകാൻ സാധ്യതയുണ്ട്. കടുംബപ്പേര്, ഭൂമിശാസ്ത്രം, ഭാഷ അല്ലെങ്കിൽ തൊഴിൽ പോലുള്ള ഗുണവിശേഷങ്ങൾ വംശത്തിനോ ജാതിക്കോ വേണ്ടിയുള്ള പ്രതിരൂപങ്ങളായി വർത്തിച്ചേക്കാം. ജാതിയെ അടിസ്ഥാനമാക്കിയുള്ള വിവേചനം, സമത്വം, അന്തസ്സ്, സാമൂഹിക നീതി എന്നിവയുടെ ഭരണഘടനാ മൂല്യങ്ങൾക്ക് എതിരാണ്, ഇത് തടയാൻ ധർമ്മികമായി അത് ആവശ്യമാണ്. വംശീയവും

ലിംഗപരവുമായ വിവേചനത്തിന്റെ സ്വാധീനമുള്ള ഒരു മാതൃക, അത്തരം വിവേചനത്തിന് കഴിയില്ലെന്ന് വാദിക്കുന്നു തികച്ചും മത്സരാധിഷ്ഠിതമായ ഒരു അന്തരീക്ഷത്തിലാണ് ഇത് സംഭവിക്കുന്നത്, കാരണം ഒരു കമ്പനി പ്രത്യേക തൊഴിലാളികളെ നിയമിക്കാനോ ചില ഉപഭോക്താക്കളെ സേവിക്കാനോ വിസമ്മതിക്കുകയാണെങ്കിൽ വിൽപ്പന നഷ്ടപ്പെടുകയും ഏറ്റവും ഉൽപ്പാദനക്ഷമതയുള്ള ജീവനക്കാരെ നിയമിക്കുന്നതിൽ പരാജയപ്പെടുകയും അത് ഒടുവിൽ പാപ്പരത്തത്തിലേക്ക് നയിക്കുകയും ചെയ്യും. തൽഫലമായി, നിരവധി സാമ്പത്തിക വിദഗ്ദ്ധർ ഇപ്പോഴും വിശ്വസിക്കുന്നത് വിപണി മത്സരശേഷി വർദ്ധിപ്പിക്കുന്നത് വിവേചന വിരുദ്ധ നിയമനിർമ്മാണത്തിന് സാധ്യമായ ഒരു ബദലാണെന്ന് വിശ്വസിക്കുന്നു. പാർശ്വവൽക്കരിക്കപ്പെട്ട വിഭാഗങ്ങൾ നേരിടുന്ന പ്രത്യേക തടസ്സങ്ങൾ പോലുള്ള യഥാർത്ഥ സാഹചര്യമായി ബന്ധപ്പെട്ട ഏതെങ്കിലും പ്രസക്തമായ വിവരങ്ങൾ ഉപേക്ഷിക്കുകയാണെങ്കിൽ, അത് മാതൃകയ്ക്കുള്ളിൽ പ്രവർത്തിക്കുന്നു.

നിങ്ങളുടെ തലച്ചോറിനെ തകർക്കൂ

1 പഠന അൽഗോരിതങ്ങൾ ധാർമ്മിക പക്ഷപാതം നേടുന്നതിന്റെ പ്രാഥമിക കാരണം:

എ. ഉയർന്ന കമ്പ്യൂട്ടേഷണൽ പവർ

ബി. ചരിത്രപരമായി പക്ഷപാതപരമായ ഡാറ്റയിൽ നിന്ന് പഠിക്കുന്നത്

സി. അമിതമായ മോഡൽ കൃത്യത

ഡി. ഓട്ടോമേഷന്റെ അഭാവം

2 താഴെ പറയുന്നവയിൽ ഏതാണ് അൽഗോരിതമിക് ബയസ് നേരിട്ട് കണ്ടെത്താൻ സഹായിക്കുന്നത്?

A. ഡാറ്റാസെറ്റ് വലുപ്പം വർദ്ധിപ്പിക്കൽ

ബി. ബയസ് ആൻഡ് ഫെയർനെസ് ഓഡിറ്റുകൾ

സി. വേഗതയേറിയ പ്രോസസ്സറുകൾ

ഡി. എൻക്രിപ്ഷൻ

3 ഒരു മോഡലിൽ ന്യായബോധ നിയന്ത്രണങ്ങൾ ഉൾപ്പെടുത്തുന്നത് ഇവയുമായി ബന്ധപ്പെട്ടിരിക്കുന്നു:

എ. ഡാറ്റ സംഭരണം

ബി. നൈതിക ഭരണം

സി. അൽഗോരിതം ഡിസൈൻ

D. ഉപയോക്തൃ ഫീഡ്ബാക്ക്

4 വിന്യാസത്തിന് ശേഷം ഉയർന്നുവരുന്ന പക്ഷപാതം ഇനിപ്പറയുന്നവയിലൂടെയാണ് ഏറ്റവും നന്നായി പരിഹരിക്കുന്നത്:

എ. ഒറ്റത്തവണ പരിശീലനം

ബി. എഡ്ജ് കേസുകൾ അവഗണിക്കുന്നു

സി. തുടർച്ചയായ നിരീക്ഷണവും അപ്ഡേറ്റുകളും

ഡി. സിസ്റ്റം നീക്കം ചെയ്യുന്നു

5 കൃത്യതയ്ക്കായി മാത്രം AI സിസ്റ്റങ്ങൾ ഒപ്റ്റിമൈസ് ചെയ്യുന്നത് ഇനിപ്പറയുന്നതിലേക്ക് നയിച്ചേക്കാം:

എ. മികച്ച ഫലങ്ങൾ

ബി. കുറഞ്ഞ ബയസ്

സി. നിലവിലുള്ള സാമൂഹിക പക്ഷപാതങ്ങളുടെ വർദ്ധനവ്

ഡി. പൂർണ്ണ നിഷ്പക്ഷത

നിങ്ങളുടെ കഴിവുകൾ പുറത്തെടുക്കൂ

1. പഠന അൽഗോരിതങ്ങളിൽ ധർമ്മികമായി അഭികാമ്യമല്ലാത്ത പക്ഷപാതങ്ങൾ ഉണ്ടാകുന്നത് ഒഴിവാക്കാൻ എന്തെല്ലാം നടപടികൾ സ്വീകരിക്കാം? പ്രഭാഷണത്തിൽ ഏർപ്പെടുക.

2. കൃത്രിമബുദ്ധി സംവിധാനങ്ങളിലെ അൽഗോരിതമിക് ബയസിന്റെ ഉത്ഭവം വ്യക്തമാക്കുക. 3. മെഷീൻ ലേണിംഗിൽ നൈതിക മാനദണ്ഡങ്ങളുടെയും തുല്യതയുടെയും പ്രാധാന്യം പരിശോധിക്കുക.